A Multi-omic Precision Oncology Pipeline to Elucidate Mechanistic Determinants of Cancer

Sunny Jewel Jones

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

Abstract

A Multi-omic Precision Oncology Pipeline to Elucidate Mechanistic Determinants of Cancer

Sunny Jewel Jones

Despite decades of effort, the mechanistic underpinnings of many cancers remain unsolved It has increasingly become appreciated that cancers can be more readily classified by their transcriptional identities rather than by genomics alone. A fuller understanding of the mechanistic connections between the aberrant genomics leading to the transcriptional dysregulation of tumors is key to both improving our knowledge of cancer biology as well as developing more precise and effective therapeutics. This thesis explores the development and application of a network based multi-omic master regulator framework designed to elucidate these pathways. In Chapter 2 we apply this analysis across 20 tumor types from the Cancer Genome Atlas and in doing so identify 407 key master regulators responsible for canalizing a high percentage of the driver genetics present across these samples. Further evaluation of these key regulators revealed a highly modular structure, indicating that the regulators work in coordinated groups to implement a variety of key cancer hallmarks. Genetic and pharmacological validation assays confirmed the predicted interactions and biological phenotypes. Chapter 3 focuses on the application of this analytical framework specifically on gastroesophageal tumors. Using a more fine-grained approach we find 15 distinct subtypes across a cohort of these heterogenous tumors. These subtypes align well with previously identified features of these cancers but also reveal novel genomic associations and key master regulators that can serve as potential avenues for therapeutic treatment.

Table of Contents

List of Figures

Main text figures:

Appendix figures:

Acknowledgments

I would first and foremost like to thank my mentor, Dr. Andrea Califano, for taking a chance on me and giving me the space, time and support to become the scientist that I am today. I came to him with no background in the field but a lot of hard-headed enthusiasm and he somehow saw some glimmer of potential in me (albeit with some reasonable reservations), and for that I am extremely grateful.

I would also like to thank my committee members, Dr. Peter Sims, Dr. Saeed Tavazoie, and Dr. Timothy Wang for their advice and insight over the years. I would particularly like to extend my gratitude to Dr. Peter Sims who was hugely helpful during my transfer to the Systems Biology department, and helped me to find my way during the early steps of the process. I'd also like thank Dr. Adam Bass for his insights over the years through our collaboration on the gastroesophageal tumor project.

While I have gotten support from a number of people during my time in the Califano lab I would especially like to thank Evan Paull for his mentorship and support and allowing me to ask him and all questions as I figured out what the heck I was doing.

In addition to my scientific mentors, I also want to thank my parents for their unconditional love and support throughout my many harebrained decisions over the years, I truly could not have gotten here without them. I'd also like to thank the rest of my family and many friends who've been supportive of me throughout this journey.

Chapter 1: Introduction

Cancer is a disease marked by uncontrollable growth of abnormal cells in the body. Over time, normal cells accrue genetic and epigenetic alterations, eventually leading to dysregulation of their physiologic behavioral patterns. These changes can come from a variety of sources, from pre-existing variants that predispose individuals to developing a certain form of cancer, to defects introduced by the machinery responsible for cell replication that are not caught before the next cell division. Most often, however, they are driven by external carcinogenic sources in our environment—including viruses, chemicals, radiation, and pollutants—leading to deleterious modifications of our DNA that changes the way cells function and interact with their environment[1].

According to the GLOBOCAN 2020 estimates from the International Agency for Research on Cancer, 19.3 million new cancer cases were diagnosed across the world in 2020 along with almost 10 million deaths[2]. By 2040 the global cancer burden is expected to reach 28.4 million cases, an almost 47% rise from 2020[3]. In the United States alone it is estimated that 39.5% of people will be diagnosed with cancer at some point in their lifetime[4]. Since the 1990s mortality rates for cancer in the US have been decreasing, with the overall cancer death rate from 1991 to 2016 dropping by a total of 27%[5]. While this is good evidence of improved early screening and treatments, there is still much we don't understand about cancer and how to treat it in the most targeted and efficient manner.

The standards of care for most cancer patients fall into three main categories: surgery, radiation therapy and chemotherapy[6,7]. Some solid malignancies, especially those caught before they disseminate and progress to metastatic disease can be surgically removed, with no additional treatment requirements. Tumors that arise in surgically inaccessible places or that have already metastasized by the time of identification have traditionally required a combination of more

aggressive treatments, including radio and chemotherapy. These broad-spectrum cytotoxic approaches preferentially target cells that are rapidly dividing—thus being effective against fast-growing cancer cells; yet, they also target healthy cells with high turnover rates, leading to significant side effects, such as hair loss (from loss of cells in hair follicles), nausea (from loss of cells along the intestine), neutropenia (from loss of white blood cells), and overall fatigue. Despite their pervasive use, a 2004 systematic review of survival rates of 22 major adult malignancies estimated that the overall contribution of chemotherapy to the 5-year survival rate was only 2.1% in the USA[8]. It's also estimated that, on average, any specific class of cancer drug is ineffective in 75% of treated patients[9,10]. Moreover, most tumors become chemoresistant after an initial response rendering further chemotherapy effectively useless. While in some forms of cancer, chemotherapy can be fully curative it is becoming increasingly clear that it may not be the best tool for most patients.

This introduction reviews the changing views of cancer biology over the years, as well as how the progression network biology and high throughput sequencing brought the field of precision oncology into a new era. It will then describe tools developed throughout the years in the Califano lab built to assess the role of master regulators as key controllers at the center of cancer cell architecture. Finally, it will then review the merging of these tools into a multi-omic framework for studying the mechanistic determinants of cancer as well as comparing it to other methods in the field.

1.1 Dissecting the Complexity of Cancer to Inform Treatment

1.1.1 Oncogene Addiction Theory

Though the general properties of cancer evolution have been known for many years the exact number of mutational events necessary to induce tumorigenesis has remained elusive[11,12]. For a while the prevailing theory has been that, in most cancers, tumorigenesis can be explained based on two to eight "driver gene" mutations[13]. These genomic mutations (here collectively referring to single point mutations, short indels, large deletions and amplifications, and fusion events) confer selective advantage to the affected cells, allowing them to proliferate and progress into tumors.

These types of genes typically fall into two categories: oncogenes and tumor suppressors. Proto-oncogenes are genes that are typically involved in cell cycle processes and regulate cell growth and division; thus, when mutated, they can produce cells with unconstrained proliferation patterns. Once these genes are mutated in a deleterious way they are considered oncogenes. A common analogy is to compare them to the gas pedal of a car getting stuck in the accelerate position. Frequently recurrent oncogenes include, for instance, *MYC, KRAS,* and *BRC-ABL*, which have been found to be mutated across a number of different tumor types and validated extensively in lab studies[14–16]. Tumor suppressor genes on the other hand, are genes that typically monitor cells for abnormal genomic events, including mutations and chromosomal breaks, and facilitate cell death or senescence when appropriate in order to maintain healthy cells and tissues. When these genes are affected by loss of function mutations—most often via truncations or deletions—they can no longer exert their regulatory surveillance within the cells, thus allowing for unchecked cell growth and division and, more importantly, rapid accrual of additional mutational events. Conceptually, they represent the brakes of a car, which if removed leave the car with no ability to

slow itself down. The most recurrently mutated tumor suppressors include *TP53*, *RB1* and *PTEN*, for instance, and many of these too have been extensively studied and validated[17–19].

In certain tumor types, the activities of a few key driver genes have been shown to effectively explain this tumorigenic process. Work by Bert Vogelstein, a pioneer in the field of cancer genetics, illuminated this progression in colorectal cancer, thus laying the groundwork for this theory of somatic evolution of cancer. Across years of research his team revealed a three-step progression that was able to transform intestinal epithelial cells into a carcinoma with metastatic potential. First, early mutations in the *APC* tumor suppressor allow cells to begin to outcompete their neighbors, in order to form a small, slow growing adenoma, and to accrue additional mutations. A follow up mutation in *KRAS* then facilitates a second round of expansion driven by the cells that now have both mutations. Later mutations in genes like *PIK3CA, SMAD4* and *TP53* promote the full development of a malignant carcinoma that can protrude out of the epithelial wall and metastasize throughout the body[13,20].

Prior to large sequencing efforts, a number of these driver genes were identified by studying familial cancers and cancer cell lines in the lab. However, as sequencing technology improved and early cohorts of tumors began to be sequenced, statistically significant patterns of mutational events started to emerge that illuminated key aspects of tumorigenesis. One example is the delineation of "mountain" vs "hill" genes. Mountain genes were those that had very high rates of mutation while hill genes were ones that occurred at comparably lower frequencies. In the initial studies of breast and colorectal cancer that outlined this mountain-hill gene dichotomy, only 5 genes were found to be mountains while about 200 were identified as hills[21,22]. These publications were some of the first of their kind to use the results of full genomic sequencing to describe the so-called landscape of cancer in the hopes of identifying new genes to both understand

tumorigenesis and ideally stratify patients for care. Though at the time all 5 of the genes identified as mountains were previously known drivers, it opened up the possibility that perhaps with more samples and more data we'd be able to reveal novel gene drivers and untangle the complexity of pathways being altered by combinations of less frequent mutant genes.

1.1.2 Precision Oncology

The goal of precision medicine is to utilize patient specific biological information in order to select and prioritize targeted treatments. This aims to improve upon the standard of care in which location and stage more predominantly determine the course of treatments. Though it's been known for a while that these methods of treatment lead to differential response, it's only been with the technological advances in next generation sequencing (NGS) and other omics technologies that the scientific community has been able to start readily identifying the biology driving these differences along with concordant diagnostics. Early successes using mutational analyses and driven by the oncogene addiction theory of cancer, led to the discovery of a number of genes, such as *BRCA1/2, TP53* and *MLH1/MSH2/MSH6*, that when mutated significantly increase cancer risk[23]. This led to a number of genes/biomarkers being used in clinical settings as companion diagnostics to facilitate tumor characterization and to prioritize treatments. A meta-analysis of 346 Phase 1 clinical trials from 2011-2013 showed that biomarker-based selection strategies were associated with both improved response rate and progression free survival, but also found that many of the studies in this time period didn't make use of these strategies[24]. While this is promising, unfortunately using mutational profiles to inform biomarkers alone has not been shown to be effectively predictive across all tumor types or across a high percentage of patients. Indeed,

a study across 21 different cancers showed that most druggable driver mutations were only present in 2%-20% of patients per cancer type, showing the limits to this therapeutic avenue[25].

In addition to oncogene targeting therapies, immunotherapies have arisen as another arm of the precision medicine field. They comprise a new class of biological therapeutics that exploit and empower the patient's immune system to fight the cancer. The idea behind these therapies is that the body already produces a variety of cell types that are geared to destroy aberrant cancer cells upon detection but that tumors evolve strategies over time to evade them by random accumulation of mutation and creating an immunosuppressive environment[26]. By tapping into each individual's immune system to promote more precise targeting of the tumors, these therapies are less globally deleterious to normal cells, though in some cases they can trigger serious and potentially lethal autoimmune reactions. Some of the foremost immunotherapy approaches include immunomodulatory antibodies, checkpoint inhibitors, vaccine immunotherapy and CAR T cell therapy[27]. Immunomodulatory antibodies work by targeting tumor-associated antigens specifically and triggering an antibody mediated immune response, or can be directly conjugated to cytotoxic drugs[28]. An extension of these are immune checkpoint inhibitors that target the receptor pathways used by cancer cells to facilitate evasion of anti-tumor T cells[29]. Cancer vaccines and CAR T-cell therapy work by stimulating T cells with tumor antigens directly, either *in vivo* or *ex vivo* respectively, in order to stimulate a tumor specific immune response[30]. While these therapies have been hugely effective in some cases, particularly in combination with traditional interventions and other biomarker therapies, this has only been the case in a minority of patients and in specific tumor types. This is in part due to the immune landscape heterogeneity across patients as well the complex mechanisms underpinning the interactions between cancers and the immune system[28,31].

While some of the above strategies have proven to be effective in a subset of cases and certainly moved the field forward, much work is still left to be done to realize the full the goal of precision medicine for cancer.

1.2 The Sequencing Era of Cancer Genomics

1.2.1 TCGA Implications and Limitations

The Cancer Genome Atlas (TCGA) initiative started in 2006 with the ambitious objective of producing a comprehensive molecular characterization of virtually all the most frequent human malignancies, including systematic profiling and characterization of the genomic alterations underlying them. As sequencing costs decreased dramatically over the last 20 years, this objective became realistic on a massive scale. After successfully completing a pilot for three tumor types (lung, ovarian and glioblastoma) the project was renewed and went on to profile more than 11,000 samples, across 33 tumor types, ultimately generating > 2.5 petabytes of data[32]. Multiple, complementary omics modalities were acquired from each tumor sample, including mutational, copy number, gene expression, methylation, microRNA and proteomics profiles, as well as clinical data about the patients from which the samples came, often with complementary technologies (e.g., gene expression microarrays and RNA-Seq). This data has resulted in numerous publications and uncovered a multitude of aspects of cancer biology that were previously too rare or disparate to be appreciated.

Yet, for all of the questions it answered, the TCGA effort failed to live up to one of its big initial expectations: identifying and classifying all tumors based on "driver gene" mutations. Indeed, most efforts to stratify tumors based on genetic alterations have produced at best mixed results, despite significant focus by many leading labs and consortia and the development of

sophisticated bioinformatic tools and pipelines[33–37]. Work by Bailey et al. in 2018 systematically merged the results of these often-divergent algorithmic efforts to create a master list of 299 "driver" genes that occurred with statistically significant frequency across multiple cohorts in the TCGA[33]. Yet, these mutations did not provide any consistent stratification of tumor samples across most malignancies. Indeed, for most tumors, overwhelming genetic heterogeneity prevented identification of genetically-distinct tumor subtypes or—with few notable exceptions accounting for only a small fraction of all tumors—subtypes presenting sensitivity to specific pharmacological agents. Instead, what has become increasingly clear is that the vast majority of tumors cannot be described just in terms of a few, highly penetrant genetic events that induce tumorigenesis in normal cells.

Part of the reason for the limited success of these taxonomical efforts based on cancer genetics can be found in the huge amount of heterogeneity that exists both between tumors (intertumor heterogeneity) and within individual cells within the same tumor (intratumor heterogeneity). Indeed, the same mutation arising in different contexts can play dramatically different roles. Cell-intrinsic properties establish a preliminary baseline, restricting the mutated proteins to interactions within pathways that are biologically available[38]. *BRAF* mutants, as an example, are found in both colorectal cancer and melanoma cells but drugs targeting *BRAF* are more effective in melanoma as compared to colorectal patients. This is likely attributable to the upregulated feedback loop with *EGFR* that occurs in colorectal tumors but not in melanoma tumors which are derived from cells with lower basal *EGFR* expression[38,39].

An additional confounding element to the identification of driver mutants is that at the time of sequencing, all of a tumor's variant genes are compiled into a uniform list, thus obscuring the critical element of time-dependent mutational order. Though a number of tools exist to reconstruct

mutational timelines from bulk sequencing data, this represents a huge resolution loss when it comes to mapping out a stepwise mutational progression as postulated by Vogelstein and others[40]. Moreover, it seems that this stepwise progression might be less prevalent than previously predicted. In colorectal cancer for instance, the poster child for this phenomenon, the frequency of tumors that have all three of the tumorigenic mutations (*APC, KRAS* and *TP53*) is lower than the frequencies of those that have none of these mutations[41]. Moreover, in certain tissues seemingly normal cells have been found to harbor just as many "driver" mutants as cancerous cells which again seems to contradict the idea that the presence of these mutations alone promotes cancer development[42]. This also complicates the driver versus passenger idea of mutations given that normal cells can harbor "driver" mutations while they are seemingly non-functional or passengers. This suggests that many mutations with weak phenotypic effects on an individual basis, and thus previously identified as passenger events, may in fact cooperate to form a strong "field effect," including private mutations that may exist only in an individual sample and cannot be identified as recurrent across a cohort. Thus, in retrospective, the potentially myopic view of focusing only on mountain vs. hill genes, while completely ignoring non-recurrent/private events, may have limited our ability to elucidate more universal mechanisms involved in tumorigenesis and cancer progression.

1.2.2 Cellular Homeostasis and Cancer

Paradoxically, while cancer's mutational landscapes are extremely heterogenous, transcriptional profiles are remarkably conserved across a large number of samples, thus producing only a limited number of distinct tumor subtypes[43,44]. Comparisons of transcriptional profiles of both normal and cancer tissue show that not only do the cancer samples differentiate from the normal tissue but that

the cancer samples themselves exhibit internal patterns of similarity, suggesting that they collectively settle into a stable state or set of states. The stability of these profiles seems to indicate that tumors rely on critical homeostatic-control machinery to maintain their transcriptional state, independent of the virtually infinite variety of genomic alteration patterns that are identified in a transcriptionally homogenous subtype.

Homeostasis is defined as the ability of a living system to preserve its steady-state operating conditions independent of the exogenous and endogenous perturbations that it is subjected to. For instance, cells are able to operate over a wide range of temperatures, even though individual biochemical reactions have much stricter temperature tolerances. Similarly, homeostatic regulation allows cells with millions of genetic differences at polymorphic sites to perform their function virtually unaffected. Indeed, this concept is important across all aspects of biology, particularly in development as cells progress from being early stem cells, which have infinite potential states, to increasingly differentiated cells that are both energetically and physiologically viable; a process called Waddington's canalization[45]. It's generally been thought that the disruption of cellular homeostasis is a key component of diseases, including cancer. While it is true that cancer cells are no longer in a state of "healthy," i.e., physiologic homeostasis, their transcriptional profiles are as stably maintained as those of physiologic cells. This suggests the existence of a dysregulated form of homeostatic control mechanisms, that has been dubbed "dystasis." Indeed, studies have shown that cells lines and xenografts derived from tumor biopsies maintain a relatively high-fidelity transcriptional state compared to that of the original specimen, even after many passages, mutational drifts, clonal selection and changes in environment[43,46,47].

1.2.3 Network Models of Cancer

Reconstructing the networks of cells to understand both the mechanisms driving their homeostasis and conversely the dysregulation occurring in cancer cells is a key strategy with which to leverage these high-throughput sequencing datasets. If a global model of all the interactions in a cell could be mapped out then biologists would in theory be able to dissect the pathways that had been disrupted during tumor development and devise precise strategies to target them. Adopting this more holistic lens, and instead focusing on the cellular networks controlling the cancer cell's transcriptional state as a way to study tumor phenotypes, introduces a far more compelling and uniform view of the biological landscape of human malignancies compared to considering only a handful of driver genes.

One way to build these networks is to use literature derived pathways that have been elucidated and constructed based on years of bench research. While these pathways can provide information that has been tested and validated a priori, these representations are both not comprehensive and lack context specificity. Moreover, these pathways, by their reductive linear nature, do not capture the complex, multivariate and dynamic systems that determine cell behavior[48]. Certainly, many important biological mechanisms have been elucidated and catalogued in databases like MSigDB, KEGG, Gene Ontology and others, and these are tools have proven to be invariably useful for looking for patterns in large datasets[49–52]. But they fall short in being able to generate truly novel network insights.

Another way to approach building these networks is inferring them directly from the data. The ability to do this is built on the assumption that interactions occurring in biological networks will generate statistical relationships in the observed data[53]. These biological interactions can be conceptualized by the "module-network" model: that genes are grouped into modules under the

same regulatory control in order to execute common functions[54]. While the increasing complexity of the network under investigation using statistical methodologies requires increasingly larger datasets to be sufficiently powerful, by pooling genes together the module-network framework increases the statistical ability to identify these structures[55].

One method to probe cellular networks and modularity directly from the data is to use methods based on SigArSearch (significant area search) which frame the search for modules as an optimization problem[56,57]. These methods first score the nodes (molecules) and/or edges (interactions) in the system based on some molecular metric, typically gene expression levels, though other measures are used as well. Subsequently a scoring function is devised to capture and quantify each aggregate subnetwork as based on the scores and activity of the member nodes and edges. From this a search strategy is deployed to identify high scoring subnetworks that thus represent an active module[56]. While these methods can work effectively in smaller networks, they become very computationally difficult as the size increases. This leads to the need for various heuristics and constraints that can keep from finding the true maximally scoring modules and networks.

Another framework to assess module structure involves using network diffusion and propagation models to integrate together other non-gene expression data types to find recurrent occurrences across a priori networks. One such method, HotNet, considers genomic alterations as heat sources, and allows them to propagate through a network of protein-protein interactions and linkages using a heat diffusion model[58]. It then computes the intensity of the heat appearing the nodes to identify relevant sub-networks. This approach is based on the idea that functionally relevant genomic events will aggregate within modules and this will be reflected in the flow of the

heat through the diffusion network. TieDIE is an extension of this that incorporated transcriptional data as well[59].

1.2.4 Gene Regulatory Network Models

Another way to frame the module-network problem is to look specifically for networks of genes under the control of transcription factors. Prioritizing the reconstruction cellular networks in this way aligns well with what we understand from developmental biology and control of cellular processes. In the context of normal cell differentiation, key transcription factors facilitate progression by coordinating expression of related sets of genes in order to implement the functions necessary for the various cell states along the cascade. Once a cell has settled into its terminal state, certain transcription factors remain active in order to maintain the state and respond to stimuli to preserve homeostasis. Moreover, scientists aiming to control these processes experimentally will promote sequential expression of various transcription factors in order to create more or less differentiated cells[60–62]. In this way, transcription factors and associated regulatory co-factors, serve as mechanistic control centers for the regulatory networks governing cells. Thus, it serves to reason that disruption of their typical activity is a key mechanistic component of tumors' dystatic network logic.

Early analyses based this framework were implemented in simple genomes like E. coli and yeast, but ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks) developed in the Califano lab in 2005, was the first to effectively reverse engineer a mammalian cellular network in a context specific fashion[63]. In order to generate these networks, ARACNe first identifies statistically significant coregulation between pairs of genes, specifically TFs/coTFs and potential targets, via mutual information in their expression levels. It then makes use of the data

processing inequality theorem to eliminate second and third order (i.e., indirect) interactions and optimally prioritize for direct regulator-target pairs. In doing so ARACNe is able to trace the most direct connection between pairs of genes thus increasing specificity of the relationships identified and reducing false positives found that can arise from indirect associations. Interestingly, in this first deployment of ARACNe it was found that the network topology was dominated by a handful of highly connected nodes that represented the majority of the connections. ARACNe has since been extensively experimentally validated and shown to be useful for reconstructing networks in a number of different tissue contexts[63–65].

A number of Bayesian based methodologies have also been employed in gene regulatory network inference as well. One such way to do this is to construct a Bayesian network to infer the influence of a particular transcription factor on any given gene[66]. Typically the nodes of the network are gene expression but other relevant experimental data and/or priors can be included as well[67]. These are then related to hidden variables that represent the influence of the relationship between the TF and target and then these hidden variables are then related to the observed quantities from the experiment, based on a joint probability model. The relative probability for each type of potential influence is then calculated and the highest one is selected. While these networks can be very effective for incorporating priors and handling uncertainty they are very computationally intensive, especially for highly complex networks, and fail to reliably predict feedback loop associations[68].

Another popular method for inferring gene regulatory networks from gene expression data is GENIE3 as developed in the Aerts lab[69]. In contrast to the previous frameworks, GENIE3 instead trains a series of random forest models to predict the expression of each gene in a dataset using the expression of the TFs as input. The models are then leveraged to derive weights for the

TFs with respect to their relevance for the prediction of expression for a particular target gene. GENIE3 also makes use of a regression element which allows for identification of non-linear relationships between regulator-target pairs. This approach has been shown to perform very well in DREAM (Dialogue on Reverse Engineering Assessment and Methods) challenges, particularly in E. coli and yeast networks, but it too is very computational expensive, particularly in complex systems.

A complementary approach that has been employed in later iterations of GENIE3, and by others, is the incorporation of cis-regulatory sequence analysis to refine predict TF-target associations. This is predicated on the fact that transcription factors typically exert their regulatory control via binding to the DNA in close proximity to their target genes. By incorporating this information, as the Aerts lab has done with later tools like iRegulon, RcisTarget, and more recently in their single cell algorithm SCENIC, they can refine their original predictions of TF-target pairs by only including interactions that also have the correct binding motif[70,71]. While this line of analysis can be useful, it is also limited by the sensitivity and range of the motif prediction algorithms and potential experimental artifacts from binding assays. It also filters out potentially crucial co-factors that may play an important regulatory role but do not bind to the DNA directly.

The above methods represent a select set of some of the key tools and frameworks that have been developed over the years to interrogate and reconstruct the networks underpinning cellular architecture. Compared to many of these frameworks, ARACNe excels in its unsupervised nature, and its limited reliance on heuristics and a priori networks. Moreover, its focus on first order interactions between regulators and their target genes provides a mechanistic lens, rather than just an associative one, to interrogate the key drivers of both individual modules as well as overall cell state.

1.3 Master Regulators at the Center of Cell Identity

1.3.1 Master Regulators

Developmental biologists coined the term "Master regulator" to refer to gene products (usually transcription factors) that are necessary and sufficient to induce cellular morphogenesis and lineage differentiation as described previously[43,72,73]. Cancer biologists later adopted the term but in a looser fashion, instead applying it to genes that were sufficient to induce a particular tumor phenotype, but not necessary *per se*. This definition encapsulates canonical oncogenes and tumor suppressors that have been shown to transform normal cells into cancerous ones. But notably for a number of these, studies have found that cancer cells without those particular mutants can still exist in the same transcriptional space, implying that these Master Regulators are not actually necessary to achieving that state. For the purposes of this paper and work done in the Califano lab, we adopt a modified, yet stricter definition of Master Regulators, requiring these proteins to be both **necessary** and **sufficient** to implement a specific cancer transcriptional state, via their mechanistic targets[43].The last part of this nomenclature is important both from a biological and statistical standpoint. Biologically, it creates a threshold that ensures selection of proteins that would be effective pharmacological targets because they are directly enacting the signature versus selecting proteins that might be important and physiologically adjacent but not critical. Statistically, prioritizing direct regulator-gene target interactions provides a stronger signal when reconstructing gene network topologies as described previously. For the purposes of this thesis I will use the term "transcriptional regulator (TR)" to signify any transcription factor (TF) or co-transcription factor (co-TF) that regulates gene expression. Among these, "Master Regulator" (MR) proteins will refer to those that have been determined to mechanistically regulate the genes that are differentially expressed in a specific cancer phenotype.

There is an inherent difficulty in determining MRs from protein activity directly in part because massive scale proteomics are still too costly for the amount of information they reveal and for this reason are less ubiquitously used. They are also dependent on the availability of high-quality, high affinity, high specificity reagents, rendering it difficult for them to be comprehensive[74]. Though more readily attainable given technological improvements, gene expression and mutational profiles of TRs are also limited in their scope. For one, while direct mutations in TRs can change their activity, this is not always the case potentially due to buffering activity of the cell to defend against these alterations. Moreover, changes in regulatory patterning have been shown to occur in the absence of these mutations. This is similarly true when looking for changes in gene expression of TRs as readouts of perturbed activity. This is largely due to the biology of how proteins, specifically TRs work. While overall expression and translation of TRs is certainly important, once present in the cell the real-time activity of TRs is usually controlled by factors like post-translational modifications, i.e. phosphorylation or methylation, sub-cellular localization, epigenetic changes, and cofactor binding availability. These elements all mechanistically contribute to whether or not a TR is able to control expression of its target genes and are difficult if not impossible to ascertain directly from proteomics assays, mutation profiles and gene expression alone.

Several tools have been developed over the years by the Califano lab to address these challenges via interrogation of regulatory networks, thus allowing direct elucidation of candidate MR proteins. Since the expression or post-translational state of a protein, in isolation, is unlikely to determine its role as a MR (as discussed above), these tools look at the targets of a TR, instead, as a readout of its activity. This is akin to using a highly multiplex gene reporter assay to measure the effect of a TR on the regulation of a specific transcriptional signature. This again hinges on the

idea that the main role of a TR is to facilitate the expression of a particular set of genes that then enact the various functions associated with the cell state of interest.

The information in the regulator-target networks reverse engineered by ARACNe, as described in the previous section, can thusly be used to predict and quantify the relative activity of a given TR[63]. If a TR is actively regulating its targets by activating or repressing their expression, this pattern will be reflected in the gene expression data. For instance, the targets activated or repressed by a TR will be over or under-expressed if the TR becomes aberrantly activated in a tumor, respectively. Conversely, if a TR's activity is unchanged, its targets will not be differentially expressed. VIPER (Virtual Inference of Protein activity by Enriched Regulon analysis) is the algorithm developed by the Califano lab to quantify this phenomenon[75]. This methodology has been extensively validated and used to reveal key MRs in a number of different contexts, both within the Califano lab and by others in the field[65,76–78].

1.3.2 Expanding the MR Framework

1.3.2.1 CINDY

Other tools have been developed over the years in the Califano lab that have aimed to build upon this MR analysis framework. The CINDy algorithm (Conditional Inference of Network Dynamics) was designed to infer regulatory dependencies between potential modulators upstream of TRs in order to identify which ones affect a TRs ability to regulate its downstream targets[79]. Understanding the upstream signaling cascades leading to TRs is important to better deconvoluting the pathways within the cell, while still focusing on the MRs. CINDy utilizes conditional mutual information between a candidate modulator, a TR and a target in order to infer the likelihood of a three-way interaction. Using this framework, it can determine if the mutual information between

the gene expression of a TR and expression of its target is affected by the expression of a particular modulator. This tool, and its predecessor MINDy (Modulator Inference of Network Dynamics) have successfully identified a number of both known and novel MR-modulator interactions including, EGRF-STAT1, HUWE1-MYC and CDK2-HMGA1[79–82].

1.3.2.2 DIGGIT

DIGGIT (Driver-gene Inference by Genetical-Genomic Information Theory) is another tool that tries to better inform MR biology by searching for genetic alterations that are associated with dysregulated MR activity[83]. Similar to eQTL (Expression Quantitative Trait Loci) analysis which aims to associate genomic variants with expression changes, DIGGIT performs aQTL (Activity Quantitative Trait Loci) analysis in order to quantitate the degree to which a particular genomic variant in a cohort corresponds with changes in activity of a TR. By prioritizing genomic alterations based on whether or not they correspond to activity changes of an TR, DIGGIT effectively differentiates likely driver genes from passengers based on a mechanistic framework rather than just a statistical one. This parses through the candidate genomic alterations in a way that captures potential drivers that are less frequent but may still have a small but important impact on specific MRs. It also builds on the idea that TRs are typically regulated post-translationally by other molecules that may develop mutations during tumor development and thus pass on aberrant signals to the TRs.

The main step of the DIGGIT algorithm is determining aQTL significance. In the original iteration, this was done using mutual information, but the updated version uses aREA (Analytical Rank-based Enrichment Analysis), an analytical derivative of GSEA (Gene Set Enrichment Analysis) that is also the underpinning of the VIPER algorithm[50,75]. Samples are ranked by

differential activity of a specific TR (after performing VIPER analysis across a cohort) and aREA is then used to calculate the relative enrichment of samples harboring a specific genomic event and having dysregulated activity of the TR. This produces a statistical determination of how likely it is that genomic event of interest affects that TRs activity. The resulting aQTL predictions can be further refined by only considering those that are also predicted to be upstream modulators by CINDY. In this way it is able to leverage both gene expression information and genomics profiles in order to improve the understanding of the mechanisms driving the activity of the MRs in a particular tumor.

1.3.2.3 PrePPI

An additional layer of information that can further bolster the ability to elucidate regulatory and signaling mechanisms is related to physical protein-protein interactions (PPIs). We know that proteins enact their various functions by physically interacting, stably or transiently, with their cognate binding partner molecules in the cell. This includes, for instance, transcription factors binding to DNA to promote transcription, kinases binding to their substrates to phosphorylate them, and motor proteins "walking" down actin filaments via structure-mediated interactions. Structural biologists have characterized many protein-protein interactions by resolving the structure of the interacting pair at the atomic level, for instance using X-ray crystallography. However, only a very small fraction of all actual interactions has been characterized in this fashion. To further expand the universe of high-likelihood PPIs, the PrePPI (Predicting Protein-Protein Interactions) algorithm was developed in collaboration with structural biologist Barry Honig's lab. PrePPI uses a Bayesian framework to integrate information from structural homologs known to physically interact, as well as from additional non-structure-related evidence, to predict the

likelihood of interaction between any two candidate proteins[84]. When assaying two novel proteins that do not have interaction information, PrePPI will search for proteins that are structurally similar (and/or have structurally similar subunits) and determine if any of these structural neighbors are part of a complex that has been reported in protein structure databases. From this PrePPI calculates a similarity score between the novel proteins and their structural neighbors and how likely it is that aspects of the neighbor interaction(s) also exist between the novel proteins. If available, non-structural evidence like co-expression or functional similarity is also incorporated into the final score via a Bayesian classifier. This algorithm has been able to generate over 300,000 high confidence interaction predictions, many of which have also been validated. This information can thus provide valuable insight to the likelihood of potential interactions between candidate modulators and MRs.

1.3.3 Oncotecture Hypothesis

Focusing on these key regulatory factors as drivers and maintainers of cancer state is the cornerstone of the "Oncotecture Hypothesis"[43]. This hypothesis posits that a handful of key master regulators (MRs), working in highly autoregulated modular structures dubbed "Tumor Checkpoints", are responsible for implementing and maintaining the dystatic state of tumor cells. This is accomplished through their coordinated regulation of genes that implement well-established cancer hallmark programs, such as proliferation, migration, immune evasion, and epithelial mesenchymal transformation (EMT). Furthermore, the Oncotecture hypothesis implies that Tumor Checkpoint proteins are likely to be downstream of relevant genomic alterations via a complex, non-linear field effect leading to their dysregulation. Given the finding that driver gene mutational profiles do not readily co-segregate with the transcriptional state of tumor subtypes, we

instead expect that different mutational patterns will induce the same transcriptional state, due to the integratory logic present in human cells. In this way these Tumor Checkpoint MRs are predicted to canalize the effects of various lower effect mutations and aberrant signals in order to implement a stable cancer cell state.

1.4 Multi-omic Analyses of Cancer

1.4.1 A Pancancer Multi-omic Master Regulator Analysis

The ARACNe and VIPER algorithms represent the conceptual foundations of MR analyses. While they have been used to elucidate mechanisms of initiation and progression in several distinct disease contexts, a systematic pan-cancer analysis across a comprehensive repertoire of tumor samples had not been performed. Elucidating both commonalities and differences in MR proteins controlling the state of different tumor subtypes and elucidating novel MR-based subtypes was thus the objective of our recently published work as detailed in Chapter 2. The goals of that analysis were two-fold: one, to test the Oncotecture Hypothesis and to determine whether tumor checkpoints could be identified on a sample by sample basis, which would integrate the genetic alterations in that sample to implement its transcriptional state, and two, to assess whether, even within tumor checkpoints, MR proteins may form smaller, highly recurrent modular structures.

MOMA (Multi-omic Master Regulator Analysis) was developed with the purpose of integrating both genetic and transcriptomic profiles toward the elucidation of tumor subtype-specific master regulators, tumor checkpoints, and their upstream genetic determinants[85]. To accomplish this goal, we leveraged all five previously described algorithms (i.e., ARACNe, VIPER, CINDy, DIGGIT and PrePPI) in a biologically-motivated and statistically robust framework to infer tumor subtype-specific MR proteins and the genomic events leading to their

dysregulation. As discussed above, one of the key tenets of this analysis is that it focuses on the mechanisms underpinning the biological role that TRs play in the cell and how they are (a) mechanistically affected by mutations in their upstream modulators and (b) mechanistically regulate their downstream targets, in turn.

In brief, the steps of MOMA are as follows: 1) VIPER activities of all TRs are calculated across a given cohort of samples using gene expression and ARACNe generated regulons, 2) Upstream modulators of each TR are inferred by CINDy 3) the effect of mutations in these modulators on the TR is predicted by the aQTL analysis component of the DIGGIT algorithm, 4) physical interactions between each TR and its upstream CINDy-inferred modulators are assessed by PrePPI and, finally, 5) all these evidence sources are integrated to generate a single MOMA score assessing the likelihood that a specific TR may be a Master Regulator protein, thus allowing ranking all TRs from the one most likely to the one least likely to be a MR, as well as the genetic events contributing to their dysregulation. This unique lens differentiates it from other network based multi-omics analyses that rely mostly on correlation or co-expression of features but not their mechanistic association.

As is discussed more thoroughly in the manuscript, we then used these MOMA scores to drive our clustering analyses across the samples in the TCGA and to determine the tumor checkpoint MRs of each subtype. In doing so we were able to identify 112 distinct tumor subtypes and to show that, in almost all of them, only a handful of MR proteins were necessary to account for the majority of functional genomic events in that sample. We were further able to identify key modules across the most frequent checkpoint MRs indicating that a number of them work in coordinated groups to implement various hallmark cancer pathways. In doing so we thus confirm both parts of the Oncotecture Hypothesis.

1.4.2 Other Multi-omic Frameworks

A number of other tools have been developed across the field using multi-omic data to characterize and classify cancer biology. A few of these were mentioned in the previous section discussing network models of cancer, but to place MOMA in the broader context of this field a selection of key ones representing some of the major classes of methods will be covered here.

1.4.2.1 Early Integration Methods

One method for approaching multi-omic clustering and classification has been to merge all the omics together into one large matrix at the outset, as a form of "early integration," followed by clustering using any number of classical methods[86]. This enables the use of previously vetted clustering algorithms but requires a number of different normalizations in order to deal with the many types of data being merged together, each with various underlying distributions. One such method is iCluster, which posits that tumor subtypes can be modeled as unobserved latent variables that can simultaneously be estimated from any number of combined omic sources[87]. After merging these omics together, it optimizes the likelihood of the observed data using an Expectation-Maximization algorithm along with a lasso-type regularization to shrink the coefficients of non-informative features. K-means clustering is then performed on this lower dimension representation of the data to identify final clustering assignments. As this method involves the generation and manipulation of one huge matrix of features, feature selection is incredibly important in mitigating computational complexity. A notable use of this algorithm was in the recent pancancer analysis across the TCGA that identified 28 different subtypes predominately dictated by cell-of-origin patterns[88].

1.4.2.2 Late Integration Methods

In contrast to early integration methods, late integration methods first perform clustering on single-omics separately, followed by an integration of these different clustering solutions[86]. This allows for tailoring of clustering algorithms best suited to each omic type prior to merging them together. Additionally, this helps to keep from any one platform dominating the others as they are all treated separately and then weighted accordingly during the consensus clustering step. One notable drawback is that signals that may be weak in each omic separately will be lost by the time they are merged with the other omics solutions. COCA (cluster-of-cluster assignments) is one such example of this method that has also been applied across tumors in the TCGA[44]. After performing per omic clustering, each sample is then encoded as a binary vector indicating its belonging to each omic's clustering solution, and all vectors for each omic are concatenated. Consensus clustering is then performed using these sample indicator matrices to identify the final clusters. The initial application of COCA across 12 tumors of the TCGA identified 11 new molecular based subtypes, but in the later work across the full set of 33 tumors iCluster (as described previously) was found to be more powerful at revealing patterns that may not be strong enough to be identified within each omic individually[44,88]

1.4.2.3 Statistical Methods

Statistical models for multi-omics clustering are based on modelling the probabilistic distributions of the underlying data. This can allow for the inclusion of biological knowledge or other priors when choosing the underlying probability functions and can allow for probabilistic assignment of samples to multiple clusters at a time[86]. iCluster is one such version of this method as it assumes the data comes from a low dimension representation that determines the membership for each

sample. Another popular framework is PARADIGM (Pathway Recognition Algorithm using Data Integration on Genomic Models), which utilizes factor graphs to integrate omics data with information about known signaling pathways and then uses this to drive clustering[89]. While this methodology can allow for flexible integration of a number of different omics it relies heavily on known interaction pathways thus limiting its ability to utilize novel connections in its clustering stage.

1.4.2.4 Deep Learning Methods

Deep learning methods make use of multi-layered neural networks to integrate together multiple sources of information and make predictions based on underlying relationships in the data. Deep learning is a new development in the field of machine learning that has shown to improve performance in a number of different contexts, particularly in image recognition and biomedical applications[90,91]. One recent method applied a deep learning tool for dimension reduction to Hepatocellular Carcinoma data from the TCGA[92]. In doing so they were able to stratify the patients into clusters corresponding to significantly different survival, though this was somewhat expected given it was a supervised analysis to find features that corresponded with survival. It is unclear how broadly applicable deep learning methods will be on biological multi-omic analyses as they usually require many samples and few features, which is opposite of current multi-omic datasets[86]. They also tend to require the tuning of a large number of parameters, which can lead to overfitting, and it can be difficult to accurately extract meaningful features from the different layers of the neural network.

Chapter 2: A Modular Master Regulator Landscape Controls Cancer Transcriptional Identity

The following is adapted from:

> Paull EO*, Aytes A*, Jones SJ*, Subramaniam PS, Giorgi FM, Douglass EF, Tagore S, Chu B, Vasciaveo A, Zheng S, Verhaak R, Abate-Shen C, Alvarez MJ, Califano A. (2021). A modular master regulator landscape controls cancer transcriptional identity. Cell, 184(2), 334-351.e20.
>
> * These authors contributed equally

Supplementary Figures and Tables can be found in Appendix A.

2.1 Summary

Despite considerable efforts, the mechanisms linking genomic alterations to the transcriptional identity of cancer cells remain elusive. Integrative genomic analysis, using a network-based approach, identified 407 Master Regulator (MR) proteins responsible for canalizing the genetics of individual samples from 20 TCGA cohorts into 112 transcriptionally-distinct tumor subtypes. MR proteins could be further organized into 24 pan-cancer modules (MRBs), each regulating key cancer hallmarks and predictive of patient outcome in multiple cohorts. Of all somatic alterations detected in each individual sample, >50% were predicted to induce aberrant MR activity, yielding insight into mechanisms linking tumor genetics and transcriptional identity and establishing non-oncogene dependencies. Genetic and pharmacological validation assays confirmed the predicted effect of upstream mutations and MR activity on downstream cellular identity and phenotype. Thus, co- analysis of mutational and gene expression profiles identified elusive subtypes and provided testable hypothesis for mechanisms mediating the effect of genetic alterations.

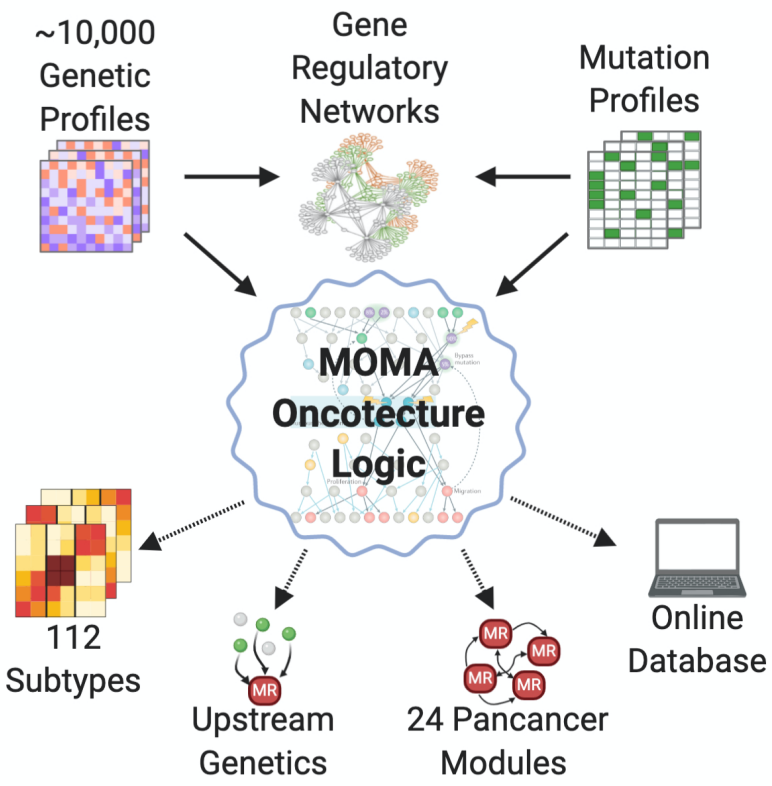


Figure 2.0 Graphical Abstract.
A high-level summary of the MOMA framework inputs and outputs.

2.2 Introduction

Our understanding of cancer as a complex system is constantly evolving: in particular, it is increasingly appreciated that the steady-state transcriptional identity of a cancer cell is tightly regulated—akin to homeostatic regulation in their physiologic counterparts—albeit via distinct and aberrant (i.e., *dystatic*) regulatory mechanisms[43]. These mechanisms play a key role in determining which transcriptional identities may be compatible with the specific set of somatic and germline variants harbored by each cell, as well as their likelihood to plastically reprogram across molecularly-distinct identities.

While some mutations effectively restrict the transcriptional identity repertoire accessible to a cancer cell—for instance, activating mutations in ESR1, FOXA1, and GATA3 are observed

almost exclusively in the luminal subtype of breast cancer[93]—many are far less deterministic. In GBM, for instance, there is only weak association between mutational and transcriptional states[94]. Despite a number of insightful studies, the molecular logic that determines the cancer cell identity as a function of its mutational and exogenous signal landscape remains elusive and largely based on statistical associations.

The *Oncotecture* hypothesis[43]—an earlier, cancer-specific equivalent of the *Omnigene* Hypothesis[95]—proposes the existence of tumor-specific Master Regulator (MR) modules (*Tumor Checkpoints*) responsible for integrating the effect of mutations and aberrant signals in upstream pathways thus determining a tumor's transcriptional identity, see Califano and Alvarez, 2017 for a recent perspective. Thus, MR analysis may help elucidate mechanisms responsible for implementing and maintaining the transcriptional identity of cancer cells, as a function of their mutational landscape, and for plastically reprogramming across distinct identities.

To study MR modularity and genetic drivers in 9,738 TCGA samples[32], on a sample-by-sample basis, we developed MOMA (Multi-Omics Master-Regulator Analysis). MOMA integrates gene expression and genomic alterations profiles to identify MR-proteins and MR-modules representing the key effectors of a tumors mutational landscape and thus responsible for implementing the cancer cell identity.

MOMA[96] can be accessed on Bioconductor[97], thus allowing analysis of virtually any cancer cohort of interest, for which patient-matched transcriptional and mutational profiles are available. In addition, the MOMA Web Application[98] provides interactive access to all results reported by this manuscript.

2.3 Results

The MOMA framework is shown in both a simplified (**Figure 2.1A-C**) and a detailed (**Figure S2.1A-E**) conceptual workflow. Briefly, *gene expression* profiles from 20 TCGA cohorts (**Table S2.1**) were first transformed to *protein activity* profiles using the VIPER algorithm[75] (Step 1, **Figure S2.1B**). Candidate MR proteins were then identified by Fisher's integration of *p*-values for (a) their VIPER-measured activity, (b) functional genetic alterations in their upstream pathways, by DIGGIT analysis[83], and (c) additional structure and literature-based evidence supporting direct protein-protein interactions between MRs and proteins harboring genetic alterations, via the PrePPI algorithm[84] (Step 2,3, **Figure S2.1C**). The vector of integrated $-\text{Log}_{10}\, p$ values (MOMA Scores) were used to weight each MR's contribution in a tumor subtype clustering step (Step 4, **Figure S2.1D**). Finally, genomic saturation analysis upstream of top candidate MRs identified those most likely to control the subtype transcriptional identity (Step 5, **Figure S2.1D**). This was followed by identification and functional characterization of MR sub-modules recurring across multiple subtypes (*MRBs*) (Step 6, **Figure S2.1E**). See Methods for a detailed description of each step.

Somatic genomic alterations considered by the analysis include single nucleotide variants/small indels (SNVs) and somatic copy number alterations (SCNAs) from the Broad TCGA Firehose pipeline, as well as fusion events (FUS) reported by PRADA[99].

VIPER has been extensively validated as an accurate methodology to measure a protein's activity, based on the enrichment of its tissue-specific activated and repressed transcriptional targets (*regulon*) in over and under-expressed genes[75]—i.e., akin to a highly-multiplexed gene-reporter assay. To generate accurate regulons for 2,506 regulatory proteins annotated as

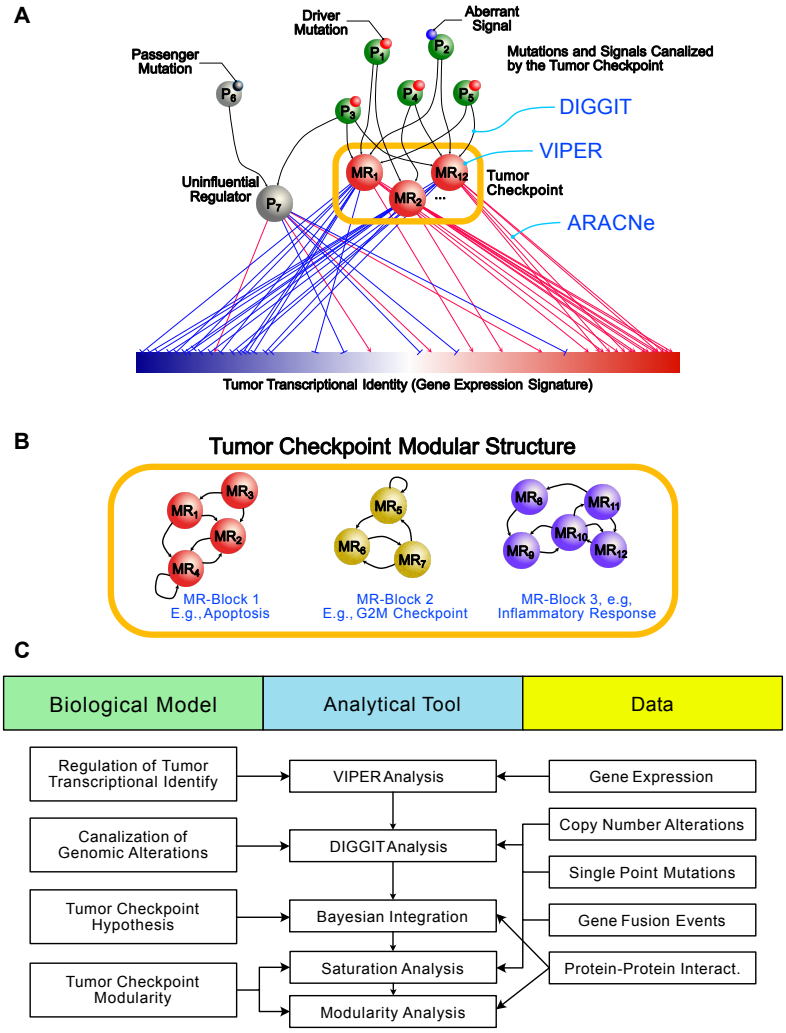


Figure 2.1 Conceptual overview of the algorithm to find sample "checkpoints" and checkpoint blocks.

(A) Conceptual diagram illustrating the "bottleneck hypothesis". Master regulator (MR) proteins (e.g., MR1 – MR12) integrate the effect of genomic alterations (small red spheres) and aberrant paracrine and endocrine signals (small blue sphere), in upstream pathway proteins (e.g., P1 – P5). Furthermore, they regulate the "downstream" transcriptional identity of the cell—shown as a gene expression signature with genes ranked from lowest (blue) to highest (red) expression—via their activated and repressed targets (red and blue edges, respectively). Passenger alterations (small black sphere) and alterations not affecting the cell's transcriptional identity occur in proteins (e.g., P6) whose downstream effectors (e.g., P7) do not affect MR activity. MR proteins form tightly autoregulated, modular structures (Tumor Checkpoints) responsible for homeostatic control of the cancer cell's transcriptional identity. **(B)** Tumor checkpoints comprise multiple sub-modular structures, termed MR-Blocks (MRBs), which regulate specific tumor hallmarks and are recurrently detected across different subtypes. As an illustrative example a tumor checkpoint comprising three different MRBs is shown. **(C)** Conceptual workflow diagram of the MOMA algorithm.

transcription factors (TFs) and co-factors (co-TFs) in Gene Ontology[52,100], we used the ARACNe algorithm[63], see Methods for more on ARACNe and VIPER accuracy.

For each candidate MR we first identified candidate upstream modulator proteins using the CINDy algorithm[79] and then assessed whether the presence of genomic alterations in their encoding genes was associated with differential MR activity (*activity quantitative trait locus* analysis, aQTL). These two steps comprise the DIGGIT algorithm, which was highly effective in elucidating key driver mutations missed by prior analyses in GBM[83].

2.3.1 Tumor Subtype identification

MOMA was used to analyze 9,738 primary samples, from 20 TCGA tumor cohorts (with $n \geq 100$ samples) (**Table S2.1**). Minimum cohort size reflected the need to generate accurate regulatory network models using the ARACNe algorithm[63]. To identify tumor subtypes representing distinct transcriptional tumor identities regulated by the same MR proteins, we performed *partitioning around medioids* clustering (PAM)[101], based on protein activity profile similarity, with each protein weighted by its cohort-specific, integrated MOMA Score (see Methods). Proteins with more functional mutations in their upstream pathways were deemed more likely determinants of tumor subtype identity and provided greater weight to the clustering solution. Within each cohort, the optimal number of clusters was determined using a Cluster Reliability Score (CRS) (**Figure 2.2A**). Using identical approaches, MR-based clustering outperformed expression-based clustering in all 20 cohorts ($p < 2.2\times10^{-16}$ in all but one cohort, SKCM, $p \leq 1.8\times10^{-8}$), by 1-tail Wilcoxon rank sum test of sample Silhouette Scores (*SS*)[102] (**Figure 2.2B**). Indeed, a majority of samples clustered by expression-based analysis had $SS \leq 0.25$—a value generally used as a threshold for statistical

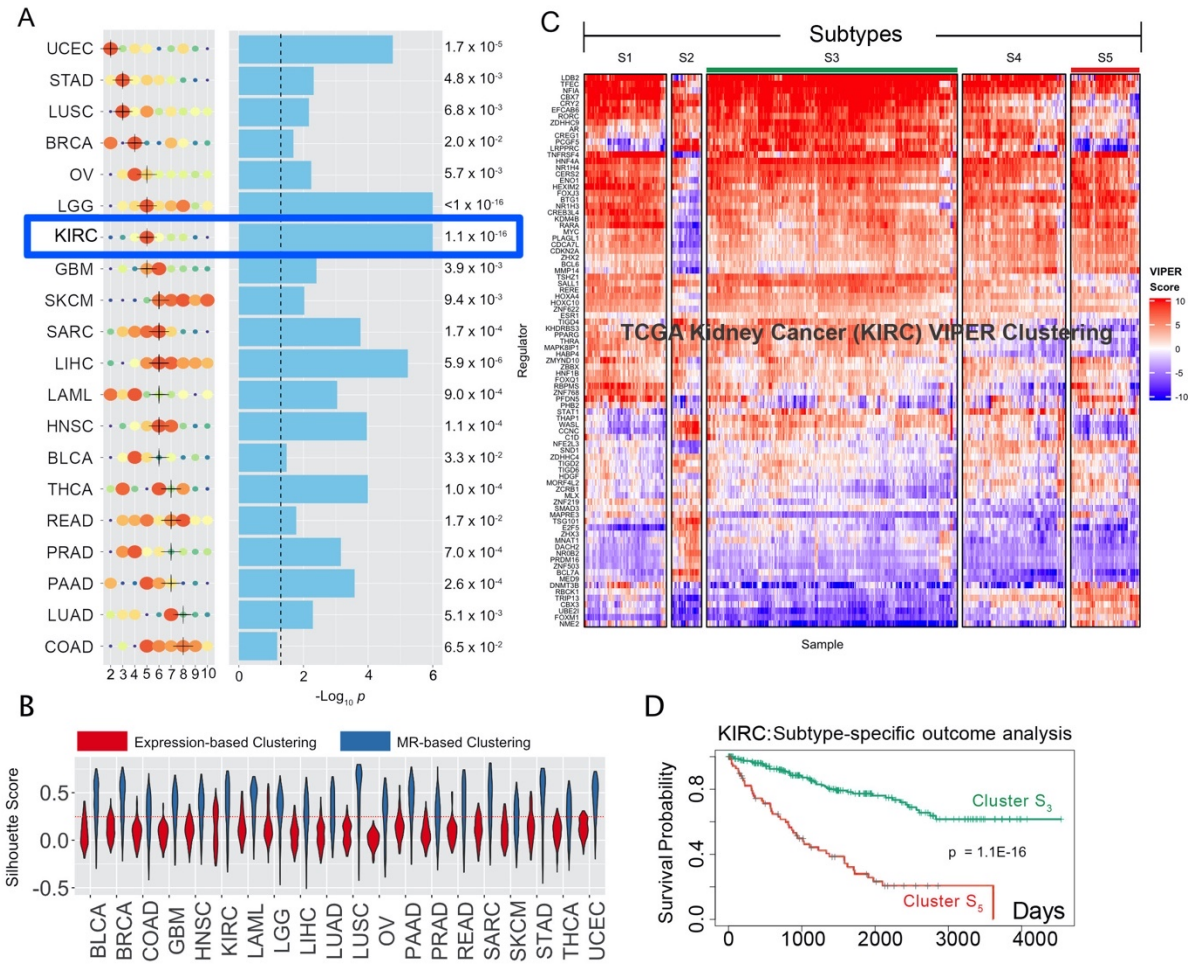


Figure 2.2 Subtypes inference by network-based integration of gene expression and mutational profile data.

(A) Cohort subtypes identified by MOMA, ranked from the lowest (UCEC) to the highest (COAD) number of optimal subtypes (x-axis). Solution optimality is shown by size and color of the dots, with larger, redder dots representing higher average CRS. The selected solution is marked by a black cross (see STAR Methods for handling ties). Statistical significance of survival separation between the best and worst clusters, by Kaplan Meier analysis, is shown next to the blue bars that represent the -Log$_{10}$ p. The dashed line represents p = 0.05. **(B)** Violin plots representing the Silhouette Score probability density (y-axis) for each of the 20 TCGA tissue types (x-axis) for the optimal clustering solution, as inferred by either MR-based (blue) or expression-based (red) cluster analysis. A dotted red line indicates the standard statistical significance threshold (SS = 0.25). **(C)** MR-based clustering heatmap for the TCGA kidney clear cell carcinoma cohort (KIRC). Rows represent Tumor Checkpoint MR proteins, while columns represent individual samples. Color scale is proportional to protein activity (red activated; blue inactivated). **(D)** Cox-proportional hazard analysis of patient survival in subtype S_5 (red line) vs. S_3 (green line) ($p = 1.1\times10^{-16}$).

significance[102]. In contrast, the vast majority of samples clustered by MR-based analysis had $SS \geq 0.25$ (**Figure 2.2B**).

Solutions ranged from $k = 2$ to 8 clusters/cohort. Whenever multiple statistically-equivalent solutions were identified, the one yielding the best survival stratification was selected (**Table S2.1**). The 5-cluster solution for Kidney Renal Clear Cell Carcinoma (KIRC) is shown as an illustrative example (**Figure 2.2C**), including differential outcome for Cluster 5 (worst) vs. Cluster 3 (best) (**Figure 2.2D**) ($p = 1.1\times10^{-16}$). Equivalent analyses for all cohorts can be accessed via the MOMA Web App, see also **Figure S2.2A** and **Table S2.1**. MOMA identified 112 subtypes, representing the stratification of cancer into transcriptional identities regulated by distinct Tumor Checkpoints (**Figures 2.2A, S2.1D**; **Table S2.1, Table S2.2**, and **Table S2.6**).

MOMA identified subtypes and differential outcome in cohorts that had been previously challenging from a gene-expression analysis perspective. For example, except for the neuroendocrine subtype, expression-based stratification of prostate cancer outcome has been elusive, requiring additional metrics (e.g. Gleason Score) or assessment of spatial tumor heterogeneity from multiple biopsies[103], which may not be available for all tumors. In contrast, MOMA identified transcriptional clusters presenting statistically significant outcome differences in 19 out of 20 cohorts (Figures 2A, S2A). Even in COAD a clear trend was detected ($p = 0.07$). Considering the significant improvement in cluster statistics (**Figure 2.2B**), this suggests that MOMA significantly outperforms expression-based subtype analysis leading to a more granular subtype structure that improves outcome stratification.

Despite its unsupervised nature, MR-based clustering recapitulated established molecular subtypes and outcome differences. In breast cancer, concordance with Luminal A, Luminal B and triple-negative subtypes was highly significant ($p = 2.2\times10^{-16}$ by χ^2 test, **Figure S2.2B**). Similarly,

in GBM, MOMA subtypes recapitulated previously published subtypes ($p = 2.2\times10^{-16}$)[104], with similar outcome stratification based on activity of established MR proteins, CEBPβ, CEBPδ, and STAT3[64] (**Figure S2.2B, S2.2C**). Best and worst survival were associated with proneural ($p = 3.0\times10^{-6}$, by Fisher's Exact Test, FET) and mesenchymal ($p = 1.3\times10^{-3}$) tumors, consistent with prior literature[64,83,104]. Virtually identical results emerged for FOXM1 and CENPF in prostate cancer, previously validated as synergistic Master Regulators of aggressive disease[105]. Prior analyses were performed by pre-selecting genes, for instance by differential expression in best vs. worst survival samples (supervised analysis), while MOMA is completely unsupervised. Notably, subtype S_6 (poorest outcome), in PRAD, comprises only nine samples—since TCGA is restricted to primary samples at diagnosis—and was thus missed by prior studies.

2.3.2 Tumor Checkpoint MRs

A Tumor Checkpoint is defined as a module with the minimum MR repertoire necessary to implement a tumor's transcriptional identity by canalizing genomic events in its upstream pathways. We thus used saturation analysis to refine the initial ranked-list of subtype-specific proteins produced by MOMA analysis to a small set of candidate MRs that optimally *account for* the subtype's genetic landscape (see Methods). By "*accounting for an alteration*" we mean that it is either harbored by the MR or by the MR's upstream modulators.

If driver mutations occurred mostly upstream of Tumor Checkpoint MRs, saturation should be achieved rapidly, with only few MRs. In contrast, if mutations were randomly distributed, saturation should be very gradual. To test this hypothesis, we considered all previously described genomic events (SNV, SCNA and FUS). To avoid over counting, we consolidated same-amplicon SCNAs upstream of MRs into single regional events, and further refined these by selecting

genomic events identified by GISTIC 2.0[106]. We then plotted the fraction of all such events predicted to be in or upstream of the top *N* candidate MRs, on a sample by sample basis—averaged over all samples in the same subtype (**Figure 2.3A**)—and defined the Tumor Checkpoint as the MRs needed to achieve a predefined *saturation threshold* in each subtype (see Methods). Finally, we identified 407 recurrent MRs (**Table S2.2**) occurring in $n \geq 4$ subtypes, a statistical threshold determined by a null hypothesis model (**Figure S2.3A**). Of these, 37 were highly recurrent, occurring in $n \geq 15$ subtypes (**Figure 2.3B**). The H3/H4 histone chaperone ASF1B emerged as the most pleiotropic MR ($n = 31$ subtypes), followed by MYBL2 ($n = 30$), JUP ($n = 29$), TOP2A ($n = 25$) and TRIP13 ($n = 25$).

Consistent with the Tumor Checkpoint hypothesis, we observed rapid genomic event saturation in all but 3 subtypes (ovarian cancer subtype S_1, S_3, and S_4). For the vast majority, saturation was achieved with very few MRs, starting at $n = 4$ for THCA subtype S_6. Overall, between 14 and 52 MRs (i.e., 0.6% to 2% of 2,506 transcriptional regulators, respectively) were sufficient to account for the first and third quantile of each sample's mutational burden, with a median of 33 MRs (1.3% of regulatory proteins). Ovarian cancer was an outlier with 170, 140, and 140 MRs needed to account for the mutations in subtypes S_1, S_3 and S_4, respectively, likely due to the very large number of likely passenger structural events in this cohort. In contrast, when MRs were chosen at random from all transcriptional regulators, saturation increased very gradually with only 0.4% of the events found upstream of 100 randomly selected MRs (**Figure 2.3A**).

At the saturation point, ~50% of all genomic events were accounted for, with a ratio of genomic events/MRs ranging from $r = 0.02$ (i.e., one event affecting 50 MRs) to $r = 32$ (i.e., 32 events affecting a single MR) and an average of 5 events/MR. This supports the role of Tumor

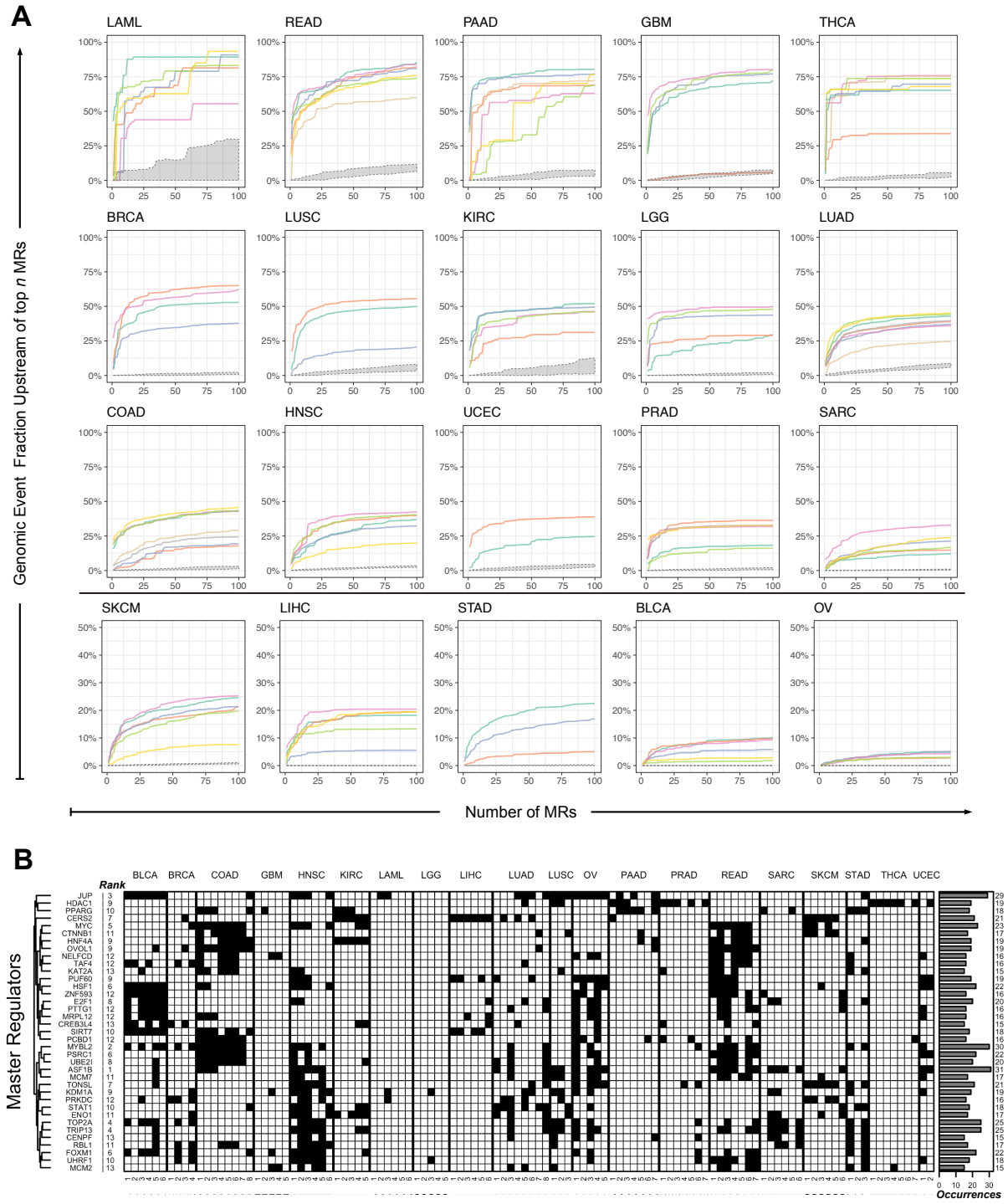


Figure 2.3 Genomic saturation analysis of candidate master regulators across all subtypes. **(A)** Individual curves show the average fraction of functional genomic events in each sample identified upstream of the top n MOMA-inferred MR proteins for each subtype, as n increases from 1 to 100. Saturation curves produced by the null-hypothesis—i.e., n randomly selected MRs from 1,253 non-statistically significant regulatory proteins (i.e., the bottom half of all MOMA-ranked proteins)—are shown in gray. Cohorts are sorted in decreasing order of the fraction of genetic events accounted for by their Tumor Checkpoint MRs. For visual clarity, the last 5 cohorts

(continued from previous page) are shown on an expanded y-axis scale (0-50%). **(B)** This panel shows the 37 most recurrently activated MR proteins, which canalize genetic alteration effects in n ≥ 15 MOMA-inferred subtypes (black cells), based on saturation analysis. Rows represent MR proteins clustered by their subtype-specific activity, to highlight MRs co-activated in the same clusters (e.g. FOXM1 and CENPF), while MOMA-inferred subtypes are shown in the columns, grouped by tumor type. The recurrence rank of each MR, based on the number of subtypes in which it is aberrantly activated, is shown to the left of the matrix while the number of subtypes is shown on the right as a bar chart.

Checkpoints as regulatory bottlenecks responsible for canalizing upstream mutations and suggests that <50% of all genomic events may be actual passengers.

To further assess MOMA's ability to differentiate between driver and passenger events, we assessed the differential enrichment of mutations upstream of MRs in either GISTIC2.0/CHASM-predicted driver events or all genomic events reported by the TCGA Firehose pipeline. When averaged across all MOMA-inferred subtypes of a specific TCGA cancer cohort, differential enrichment of the former was highly statistically significant across all but one tumor cohort (LAML), with *p*-values ranging from $p = 10^{-7}$ to $p = 10^{-156}$ and significant fold-ratio with respect to the latter (**Figure S2.3B**, **S2.3C**). This suggests that low SNV and high fusion-event rates, may have contributed to the LAML discrepancy, since CHASM only assesses candidate SNVs. Even though a majority of inferred events were previously unreported, MOMA effectively recovered all but one (RQCD1) of the 200 high-confidence pancancer driver genes reported in (Bailey et al., 2018), as well as 82.3% of the high-confidence, tumor-specific driver genes, averaged across all subtypes (min:50%, max:100%, **Table S2.3**).

In colon adenocarcinoma (COAD), for instance, 8 subtypes were identified, including 4 enriched in MSI^{High} samples (S_2, S_3, S_7, and S_8), two dominated by single nucleotide variants but not enriched in MSI^{High} samples (S_1 and S_4), and two dominated by focal SCNA events (S_5 and S_6). The mutational landscape of these subtypes was highly distinct. For instance, the classic tumor

suppressor APC was frequently mutated in all subtypes (S_2 = 39% to S_5 = 93%) except S_8. Similarly, taken together, mismatch repair genes (MSH2, MSH6, and MLH1) were mutated in ~50% of S_2 but not S_3 samples, while BRCA2 was disproportionally mutated in S_3 and several other genes were uniquely or disproportionally mutated in either subtype (**Figures 2.4A, 2.4B**). Finally, PI3K pathway mutations were frequent in S_2 and S_3, yet rarely mutated in other subtypes. In contrast S_5 and S_6 were dominated by focal SCNA events, with several genes mutated exclusively or disproportionately in S_5, while virtually all S_6 mutations were also detected in S_5 (**Figure 2.4D**). Similar mutational co-segregation differences were detected across virtually all cohort subtypes.

Regional (i.e., non-focal) SCNAs have been largely ignored by previous analyses, due to their high gene content. However, MOMA is effective at removing regional SCNA genes that are unlikely to modulate MR activity, by DIGGIT analysis. When regional SCNAs were included, subtypes became highly homogeneous in terms of their mutational repertoire across patients. Consider, for instance, COAD subtype S_5 where, except for $APC^{Mut/Del}$, already present in 98% of samples, the top 10 regional events increased in frequency from 12.5% to 84%, when focal and regional SCNAs were analyzed together (bold red, **Figure 2.4E**).

2.3.3 Tumor Checkpoints are Hyperconnected and Modular

Analysis of existing molecular interaction networks confirmed that Tumor Checkpoints represent hyperconnected modules, compared to equisized protein sets chosen at random from 2,506 regulatory proteins, as a null model. Networks include HumanNet 2.0[107] ($p < 5.0\times10^{-42}$, by Kolmogorov-Smirnov, **Figure S2.4A**), Multinet[108] ($p < 2.0\times10^{-37}$, **Figure S2.4B**), and PrePPI[84] ($p = 9.0\times10^{-44}$, **Figure S2.4C**).

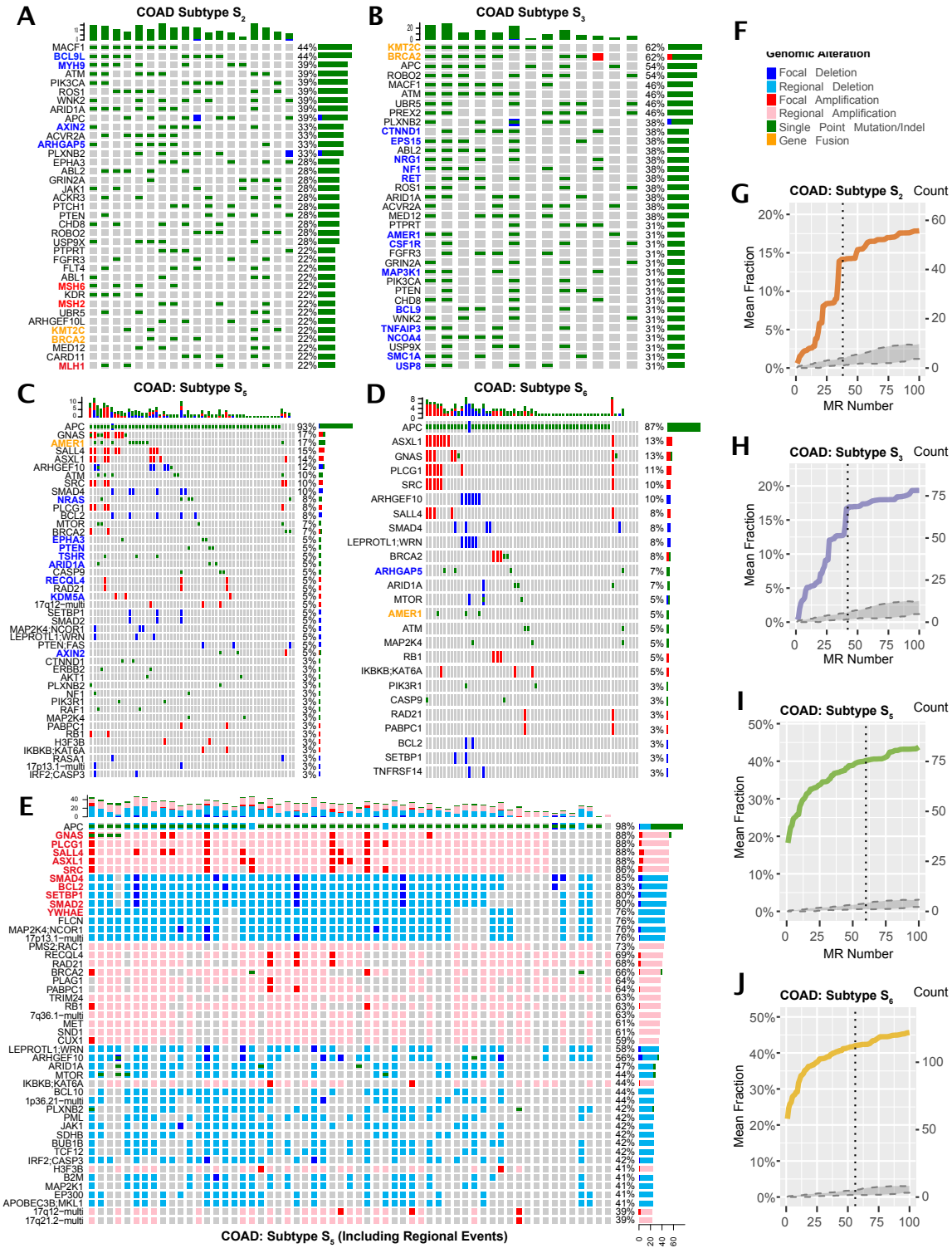


Figure 2.4 Genomic Alterations Dysregulating COAD Tumor Checkpoints.

(A – D) OncoPrint plots[109] showing genomic alterations in pathways upstream of subtypes S2/S3 (MSIHigh) and S5/S6 (MSS) in COAD. Only focal SCNA events are shown. Horizontal histograms and percent numbers show the fraction of samples harboring the specific event type. Vertical

(continued from previous page) histograms show the number of events detected in each sample. For SCNAs, each row corresponds to an independent cytoband, identified by a functionally established oncoprotein/tumor suppressor. Blue labels represent genetic alterations detected only in one subtype but not the other (i.e., S2 vs. S3 or S5 vs. S6), orange labels show alterations disproportionately represented across subtypes, while red ones show mismatch repair genes in S2. **(E)** OncoPrint plot of S5 alterations, including those in Regional (i.e., non-focal) SCNA, with most affected events shown with a red label. (F) Legend for genomic event types. (G – J) Genomic saturation curves for COAD subtypes S2, S3, S5, and S6. Vertical dashed line indicates the saturation threshold, see Figure 2.3A for detailed description.

We then tested whether subtype-specific Tumor Checkpoints may be decomposed into finer-grain MR sub-modules—recurrent across multiple subtypes—representing pancancer core-regulatory structures. Clustering of 407 MRs identified by saturation and recurrence analysis yielded 24 MR-Blocks (*MRBs*) as an optimal solution (**Figure S2.5A**), with each MR assigned to a single MRB (*core-set*). Since individual TFs may perform different functions, depending on interacting co-partners (e.g., MYC/MAX vs. MYC/MIZ-1), we used a "fuzzy" clustering algorithm to refine core-sets with additional non-unique MRs[110] (**Figures S2.5B, S2.5C ; Table S2.4**).

Each Tumor Checkpoint is thus deconstructed into a specific combination of activated or inactivated MRBs (**Figure 2.5A**), with MRB activity computed as the average activity of all of its MRs. Transcriptional targets of individual MRB MRs were enriched in Cancer Hallmarks[49,111] and KEGG/Reactome categories[51,112] (**Figures 2.5B, S2.5D; Table S2.4**). For instance, MRB:7 and 24 regulate proliferation/DNA repair and inflammation/immune response programs, respectively, and are differentially active across subtypes (**Figures 2.5A, 2.5B**). Consistently, MRB activity effectively stratified outcome in multiple datasets, see METABRIC BRCA and TCGA SKCM, for instance (**Figures 2.5C, 2.5D**). Enrichment of Tumor Hallmarks, KEGG, and Reactome categories in genes altered upstream of each MRB was generic and sparser (**Table S2.4**), suggesting that functional specificity is manifested after MRB integration, rather than in the upstream genetics.

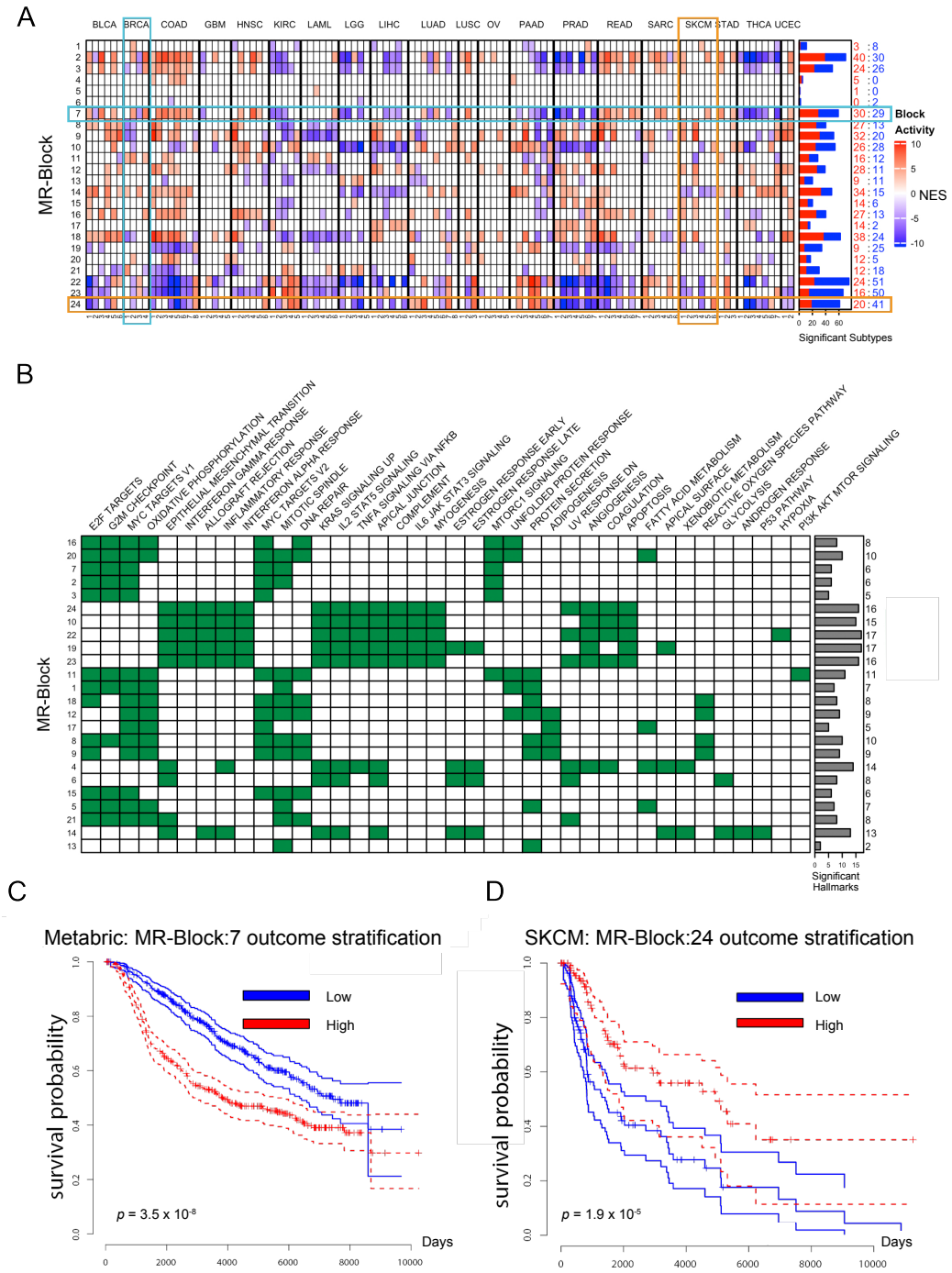


Figure 2.5 MRBs are recurrently activated in cancer and regulate established tumor hallmarks. **(A)** Heatmap showing statistically significantly activated (ON) and inactivated (OFF) MRBs for each MOMA-inferred transcriptional subtype ($p < 10^{-3}$), grouped by tumor type. Color saturation is proportional to statistical significance (Average protein activity of MRB MRs), see color-scale legend. Breast cancer (BRCA) and melanoma (SKCM) subtypes are marked to highlight differential activation of MRB:7 and 24, respectively, also highlighted. Horizontal histograms show total number of subtypes with significantly activated (red) and inactivated (blue) blocks, numerical values are also shown for clarity. **(B)** Enrichment of Tumor Hallmarks in MRB MRs

(continued from previous page) and their transcriptional targets (False Discovery Rate, FDR < 0.05, by Benjamini-Hochberg) identifies hallmarks significantly associated with each MRB. Order is based on co-clustering across both rows and columns to highlight related hallmarks and MRB co-activation. Horizontal histograms summarize the total number of enriched hallmarks per block. **(C)** MRB:7 activity stratifies survival in the Metabric breast cancer cohort ($p = 3.5\times10^{-8}$; by Kaplan Meier). **(D**) MRB:24 activity significantly stratifies survival in the TCGA melanoma cohort ($p < 1.9\times10^{-5}$). In contrast to MRB:7, higher activity of MRB:24 is associated with better outcome, consistent with its role as a marker of inflammation and immune sensing (Figure 2.5B).

2.3.4 Tumor Checkpoint MRs are Enriched in Essential Proteins

We further assessed whether the inferred Tumor Checkpoint MRs were enriched in essential proteins, based on Achilles Project data[113], see **Figure S2.5E** for a conceptual workflow. Specifically, cell lines optimally matching MOMA-inferred subtypes were identified by protein activity analysis (see Methods). Essentiality was then assessed based on Achilles' score in matched cell lines. Overall, MRs were highly enriched in essential genes ($n = 141$, $p = 7.1\times10^{-6}$; **Figure S2.5F**), based on 10^6 random selections of the same number of regulatory proteins for each subtype.

We then tested MRB-specific essentiality. As expected, those most enriched for cell viability hallmarks, such as MRB:2, 3, and 7 (**Figure 2.5B**) were most enriched in essential MRs (50%, 43.8%, and 30.4%, respectively), including proteins such as E2F1, E2F2, E2F7, TOP2A, PTTG1, FOXM1, MYBL2, UHRF1, DNMT3B, ZNF695, TCF19, RBL1, and ZNF367. Interestingly, essentiality was also prominent in other MRBs, including 31% of MRs in MRB:6 (ZNF436, HES1, HOXB7, TP63, TRIM29, GRHL1, PBX4, IKZF2, RARG, IRX5, HHEX, RUNX2, STAT5A, HDAC1, HOXC6) and 19% of those in MRB:14 (GRHL2, OVOL1, ZBTB7B), for instance. As expected, no essential MRS were found in immune-related MRBs (MRB:10, 19, 22, 23, and 24)—consistent with lack of immune function in cell lines. However, the role of many of these MRs in pancancer inflammation was previously reported[114]. This

suggests that MOMA can identify MRs that are relevant in a human tumor context but may be missed in viability assays *in vitro*.

2.3.5 MRBs Improve Outcome Analysis

To assess whether MRBs could stratify patient outcome, we used a sparse Lasso COX proportional hazards regression model[115], with MRB activities as predictors. Of the 20 TCGA cohorts, 16 could be effectively stratified, often with highly-improved *p*-values compared to Tumor Checkpoint stratification (**Figures S2.6A and S2.6B vs. S2.2A; Table S2.4**). For instance, in melanoma we observed striking survival separation ($p < 1.6\times10^{-7}$), using a 6 MRB model—including MRB:10, controlling inflammatory/immune programs (**Figure 2.5B**). Tumor Checkpoint-based analysis was much less significant ($p = 9.4\times10^{-3}$). Similarly, in colorectal cancer, significant outcome separation was achieved using a 3 MRB model ($p = 3.5\times10^{-3}$)—with MRB:6 providing the greatest contribution—while Tumor Checkpoint stratification was not significant (**Figure S2.2A**). Finally, some MRBs provide complementary stratification. For instance, MRB:6—controlling EMT, KRAS signals, and immune evasion programs—effectively stratified HNSC, GBM, COAD, BRCA, and BLCA, but not UCEC, STAD, SKCM, SARC, LUAD, LIHC, while the opposite was true for MRB:3—controlling proliferation and DNA repair programs.

To assess whether TCGA-inferred MRBs generalize to other cohorts, we analyzed the METABRIC breast cancer cohort, including metastatic samples, with long-term survival data[93]. Considering the 7 MRBs with highest differential activity in TCGA BRCA (MRB:2, 3, 7, 11, 14, 16, and 21), all of them, but MRB:11, provided significant survival stratification in METABRIC, 5 of 6 with $p < 9.1\times10^{-7}$ (Bonferroni corrected) (**Figure S2.6C**).

2.3.6 MRB:2 Canalizes Driver Mutations in Prostate Cancer

To validate the effect of genetic alterations affecting MRB activity, we selected MRB:2, the most recurrently activated across all subtypes (40/112, **Figure 2.5A**). By regularized COX regression, MRB:2 produced some of the largest outcome regression coefficients across TCGA (**Table S2.4**), emerging as one of the most significant predictors of poor outcome (**Figure S2.6A**). 11 of its 14 proteins had been previously reported as MRs of malignant prostate cancer (FOXM1 CENPF UHRF1 TIMELESS CENPK TRIP13 ASF1B E2F7 PTTG1 MYBL2 ASF1B TRIP13), including 7 out of 8 in its core-set. FOXM1 and CENPF—the 6th and 13th most recurrent MRs **(Figure 2.3B**)—were validated as synergistic MRs[105]. Yet, the mutations inducing MRB:2 aberrant activity were not previously elucidated.

MOMA identified 7 molecularly-distinct prostate adenocarcinoma subtypes, with significant survival separation (**Figure 2.6A**), including S_6 (worse) and S_1, S_3 and S_5 (best survival) ($p = 6\times10^{-3}$), as confirmed by Gleason Score and biochemical recurrence analysis (**Figures 2.6B, 2.6C**). Consistently, MRB:2 MRs are only activated in S_6 samples (**Figure 2.6A**). In addition, the S_6 vs. S_1 differential expression signature (9 and 149 samples respectively) is enriched in tumor hallmarks associated with MRB:2 (**Figure 2.6D**). We ranked MOMA-inferred alterations upstream of MRB:2 based on their statistical significance across all TCGA cohorts and selected those with the strongest MRB:2 association (**Figures 2.6E, 2.6F**), most of which were not identified as drivers by MutSig2.CV[116] and Mutation Assessor[117] (**Table S2.3**).

We selected 6 loss-of-function MRB:2-associated events for experimental validation, including TP53^{Mut} (top pancancer SNV), PTEN^{Del} and PTEN^{Mut} (top pancancer SCNA), MAP3K7^{Del} (top PRAD-specific deletion), SORBS3^{Del} (top integrated pancancer/PRAD-specific deletion) and BCAR1^{Del} (top pancancer deletion supported by MR physical interaction, with

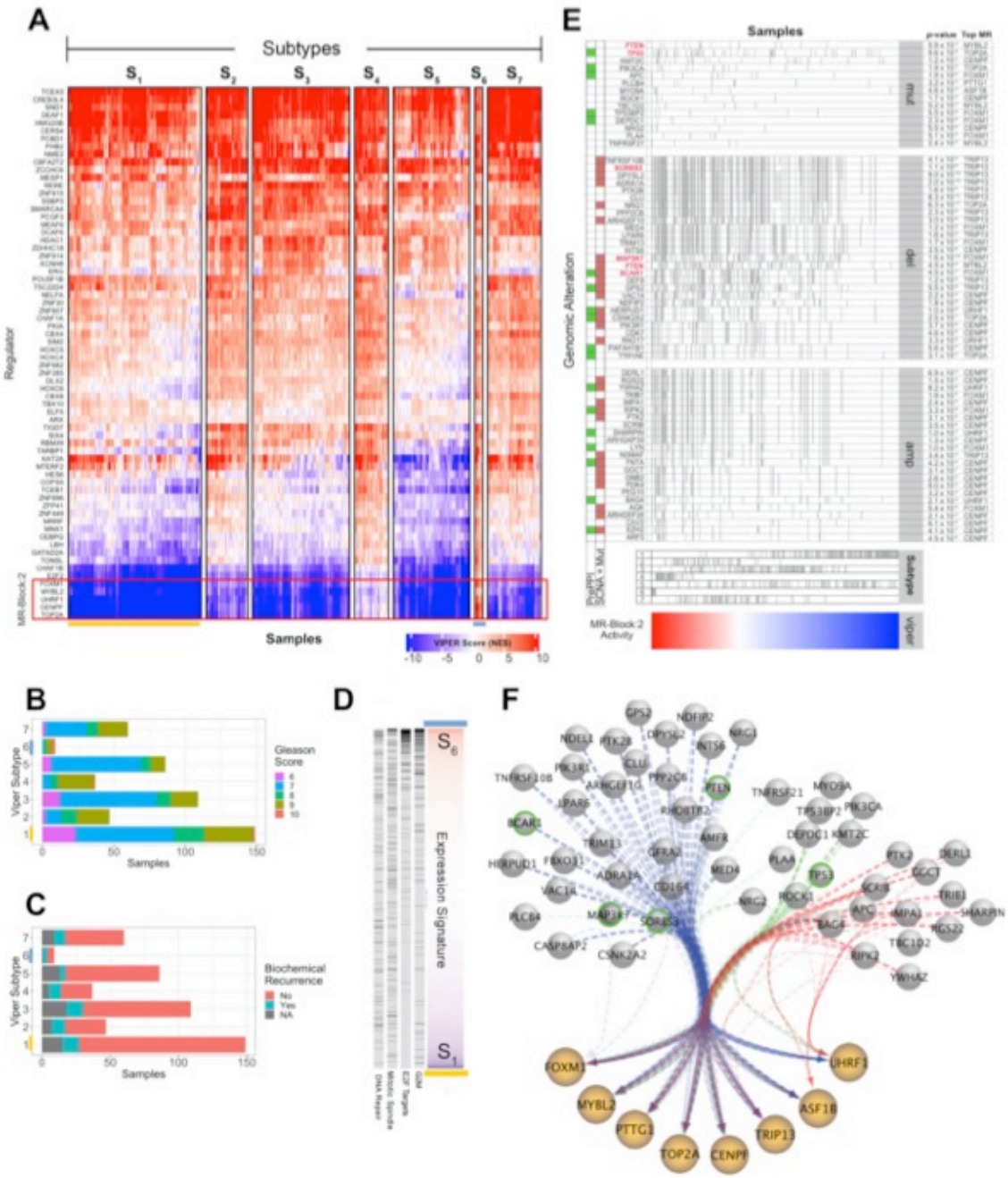


Figure 2.6 MRB2 and its upstream genetic alterations drive the most aggressive PRAD subtype **(A)** Heatmap showing MR-based clustering of the TCGA prostate cancer cohort (PRAD) into 7 molecularly-distinct subtypes, as described in Figure 2.2C. **(B)** Gleason Score frequency stratification by subtype. **(C)** Biochemical recurrence status by subtype. **(D)** Enrichment of genes in MRB:2 hallmark categories in genes differentially expressed between S_1 and S_6 subtypes, sorted by Student's t-test analysis. Genes in each hallmark are shown as black ticks and statistical significance is computed by GSEA analysis ($p < 2.2\times10^{-16}$, i.e., below minimum computable significance). **(E)** Genomic events significantly associated with MRB:2 activity. Samples (columns) are sorted by MRB:2 activity (bottom heatmap) and presence of a specific genomic event is shown

(continued from previous page) as vertical tick-marks. Functional SCNA events for genes that also harbor mutations in the cohort are marked with a brown square. Those involved in protein-protein interactions with MR proteins, based on PrePPI analysis, are marked with a green square. Events are ranked based on their subtype frequency. The top integrated aQTL, CINDy and PrePPI association *p*-value (using Fisher's method) for each event with an MRB:2 MR is shown on the right side. The five genes selected for experimental validation are highlighted in red. We also indicate the subtype designation per sample, as shown as tick marks above the heatmap. **(F)** Network diagram of MRB:2 proteins with edges representing a select set of DIGGIT-inferred alteration-MR interactions—including for deletions (blue), mutations (green), and amplification events (red)—shown as bundled edges. Green-circled events were selected for experimental follow-up.

FOXM1) (**Figure 2.6E**). Of these only PTEN, a classic prostate cancer mutation, and TP53, a hallmark of advanced, castration-resistant disease, were previously reported. We validated their functional role in 22Rv1 AR-sensitive prostate cancer cells with low MRB:2 activity, thus ideally suited to detecting activity increase in loss-of-function assays. Two shRNA hairpins/target were used. Functional and tumorigenic effects were assessed both *in vitro* and *in vivo* (**Figure 2.7A; Table S2.5**).

VIPER analysis following shRNA-mediated silencing of 4 of the 5 candidate genes vs. negative controls, revealed statistically significant activity increase of MRB:2 activity, based on its 8 core-set MRs (**Figure 2.7B**). TP53 silencing, while not significant at the MRB level, induced FOXM1, PTTG1, and UHRF1 activity increase. Functionally, MAP3K7, SORB3, PTEN and TP53 showed significant increase in cell migration, as assessed by wound healing assays at the indicated time points relative to control cells infected with scramble shRNAs (**Figures 2.7C, 2.7D, S2.7A**) This was confirmed by Boyden chamber migration assays (**Figures 2.7E, S2.7B**). Finally, 22Rv1 cells were engrafted in immune deficient mice, following target gene and negative control silencing. MAP3K7, TP53, and PTEN silencing produced significant growth increase compared to negative controls ($p < 0.01$, by two-way ANOVA) (**Figure 2.7F**).

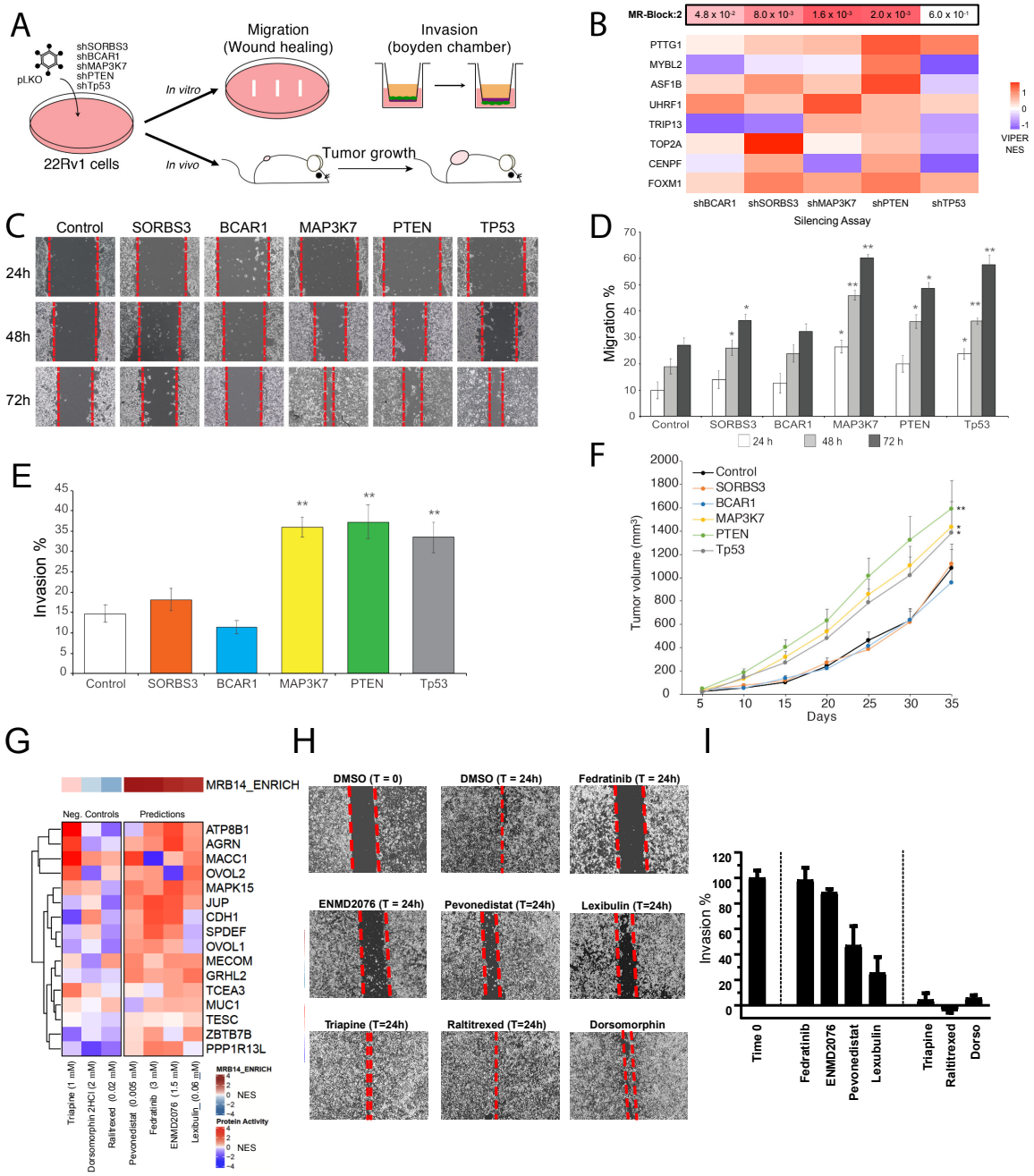


Figure 2.7 Functional validation of MRB:2 and 14

(A) Conceptual diagram of the functional validation assays. Androgen independent 22Rv1 prostate cancer cells were infected with lentiviral non-targeting control vectors and vectors containing shRNA hairpins to silence genes harboring predicted, recurrent genomic events upstream of MRB:2. Stably silenced clones were then used to perform both *in vitro* and in *in vivo* assays. **(B)** VIPER analysis of 8 MRB core-set proteins (rows) in each silencing condition (columns). Significance of overall MRB:2 differential activity is shown above. **(C)** Migration of 22Rv1 cells was assessed in wound healing assays at 24 (control), 48, and 72 hours after scratching a confluent culture of control and silenced 22Rv1, in triplicate. **(D)** Quantification of the migration

(*continued from previous page)* assay. Bars indicate the migration percentage (gap area compared to T = 24h) ± standard error of the mean (SEM). *P*-values from the two hairpins were integrated by Fisher's method (* $p < 0.05$, ** $p < 0.001$, by 1-tail Student's t-test). **(E)** Quantification of Boyden chamber invasion assays in triplicate. Bars represent the proportion of invading cells ± SEM. *P*-values from the two hairpins were integrated by Fisher's method (** $p < 0.001$, 1-tail t-test). **(F)** Functional, *in vivo* validation of tumorigenic effects. Tumor growth curves, up to 35 days, are shown for mice engrafted with control and silenced 22Rv1 cells. *In vivo* assays where performed in triplicate; * $p < 0.05$ and ** $p < 0.001$, by 2-tail, two-way ANOVA. **(G)** Heatmap showing the effect of selected drug perturbations (columns) on the activity of MRB:14 MR proteins (rows) at 24h. Drug names are followed by their EC_{20} concentration, based on dose response curves. The color bar on top of the heatmap indicates the significance of the average MRB:14 differential activity. **(H)** Modified migration assay of DU145 cells after drug treatment to activate MRB:14, assessed at 24h after drug treatment. **(I)** Average gap area (*gap remaining*) quantitation by integrating measurements of ≥ 3 images along the gap, after subtracting any residual gap area in DMSO-treated cells. Percentage gap remaining is calculated with respect to images at 0h time.

2.3.7 Pharmacological MRB Modulation

We then asked whether MRB activity and associated function may be pharmacologically modulated. We focused on MRB:14, whose activity emerged as critical in establishing and maintaining hormonally-mediated luminal epithelial identity and cell adhesion (i.e., anti-migratory) phenotypes. Several MRB:14 proteins (e.g., GRHL2 OVOL1 ZBTB7B) emerged as essential in MRB:14 active cell lines and in tissue-specific knockout mice studies[118–120]. Others—SDPEF GRHL2 JUP/γ-catenin CDH1/E-cadherin ZBTB7B OVOL1 OVOL2 ATP8B1/FIC1 PPP1R13L/iASPP—are established regulators of epithelial cell adhesion and anoikis, cellular apical-basal polarity, luminal epithelial structure maintenance, EMT, cell migration, and inflammation, as shown in prostate, breast, colon, and skin studies[121,122]. MOMA analysis recapitulated these roles in terms of hallmark enrichments, including androgen and estrogen response, EMT, apical surface and apical junction, and inflammatory response.

Consistent with our analysis, SPDEF, GRHL2, γ-catenin, and CDH1 protein expression was lost or significantly reduced in AR-insensitive (DU145 and PC-3) vs. AR-sensitive (LNCaP) cell lines (**Figure S2.7C**). LNCaP cells treated with the AR antagonist enzalutamide or DMSO[123]

confirmed that MRB:14 genes have AR-dependent expression (**Figure S2.7D**). Furthermore, their role in luminal epithelial identity maintenance was supported by luminal and basal prostate epithelial cell analysis[124] (**Figure S2.7E**). Indeed, MRB:14 activity effectively stratified luminal vs. basal samples in BRCA and BLCA TCGA cohorts, by PAM50 classification (**Figure S2.7F**), further supporting MRB:14's role as a positive determinant of hormone-signal-mediated luminal state across tissues and loss of luminal identity when inactivated.

VIPER analysis of patient-matched biopsies pre and post androgen deprivation therapy (ADT)[125] showed pronounced MRB:14 MR activity suppression (**Figure S2.7G**). Indeed, metastatic, post-ADT tumors are generally basal-like having undergone EMT, raising the question of whether prolonged ADT may induce loss of adhesion and metastatic progression[126,127]. Intermittent testosterone replacement therapy reduced appearance of aggressive tumors[128,129], reflecting potential benefit of periodic, AR-mediated cell adhesion reinforcement.

To test whether pharmacological activation of MRB:14 MRs may reduce the migratory, EMT-related potential of aggressive prostate cancer, we used the OncoTreat algorithm[130] to prioritize 120 FDA-approved and 217 late-stage (phase-II and -III) experimental drugs, based on their overall ability to activate MRB:14 MRs, using RNASeq profiles of AR-resistant DU145 cells at 24h after treatment (see Methods). Four MRB:14-activating drugs were inferred at physiologically-realistic concentrations (<10μM), including fedratinib, pevonedistat, ENMD-2076 and lexibulin (**Figure 2.7G**), and their effect was assessed in wound healing assays. All 4 drugs but none of the negative controls significantly inhibited DU145 cell migration at 24h (**Figures 2.7H, 2.7I**). The latter—triapine, raltitrexed, and dorsomorphin—were randomly selected among drugs with no significant MRB:14 activity effect (**Figure 2.7G**).

2.4 Discussion

The repertoire of transcriptional identities accessible to a cancer cell, which ultimately determine its plasticity potential, is constrained by its mutational and paracrine/endocrine signal landscape, as well as its cell-of-origin epigenetics. Yet, the specific mechanisms by which these constraints are implemented are still poorly understood. We thus attempted to establish a more direct link between the proteins that regulate a tumor's identity and the genomic alterations that induce their aberrant activity using an algorithm, MOMA, that integrates multiple omics data.

The fine-grain subtype-structure emerging from the analysis revealed a highly modular and recurrent regulatory architecture, implemented by subtype-specific, combinatorial activation or inactivation of 24 Master Regulator modules (MRBs), each regulating specific tumor hallmarks. It also highlights highly-recurrent and distinct mutational patterns within each subtype that had been missed by gene expression-based clustering. This suggests a "mutational field effect"—a term borrowed from Ising Spin Fields in ferromagnetism[131]—where many "weak" events that would be unable to dysregulate MR proteins on an individual basis—such as those in regional SCNAs—may cooperate to create a "strong" effect, as discussed for COAD. Weak event cooperativity may have been previously missed because regional SCNA contains dozen to hundreds of potential contributing genes, most of which are efficiently removed by MOMA's CINDy and aQTL analyses.

While most samples lacked a driver event quorum by conventional analyses, MOMA inferred a large number of functionally-relevant events contributing to MR dysregulation in most samples, consistent with other complex diseases[95]. Despite the remarkable complexity of these mutational patterns, our study suggests that their effect is canalized by only 112 distinct regulatory modules (Tumor Checkpoints), each representing a combination of only 24 primary MRBs.

Consistent with the notion that transcriptional cell states have emerged as more accurate predictors of drug-sensitivity, compared to genetics[132], this suggests that MR-based analyses may produce a more tractable landscape of potential therapeutic targets than could be achieved by genetic-based approaches, especially as great strides are being made to target transcriptional regulators using E3-ligases, covalent binding molecules, or antisense agents. To further support this observation, we show that MRB activity and associated phenotypes can be effectively modulated by drugs predicted to invert the activity of their MRs, suggesting that a relatively small repertoire of MRB-targeting drugs could be developed to support precision combination therapy, as determined by MRB activity on an individual patient basis.

Over the last 50 years, a number of cancer hallmarks, representing programs necessary for cancer cell survival and proliferation, have emerged[26], thus spurring research aimed at identifying the specific proteins and protein-modules that comprise them. This has led to development of several methods to 'decompose' the 20,000+ dimensional gene-expression data space into orthogonal programs, either using 2-dimentional matrices[133] or higher dimensional tensors[134], thus creating a simplified representation of the underlying cellular states and shared oncogenic alterations[133,135]. These studies are encouraging and confirm that cancer hallmarks may be indeed implemented by coordinated activity of specific gene modules. However, current hallmark representations are basically tumor-independent gene sets that lack information on what regulates or dysregulates them. MRBs provide a complementary, subtype specific representation of the proteins that causally regulate cancer hallmark gene sets and, thus, a potential way to modulate them on an individual tumor basis, as confirmed by validation of OncoTreat-predicted drugs.

MRB:2 was selected for experimental validation as the most recurrently activated across clustering solutions, mostly in poor outcome subtypes (**Figures 2.5A, S2.5C**). While 11 of its 14

proteins, which regulate cell growth, DNA repair, and mitotic programs (**Table S2.4**), were previously inferred as MRs of the most aggressive subtype of prostate cancer, including FOXM1 and CENPF validated as synergistic MRs[105], their concerted, pancancer role had been missed. Among them, TRIP13 also plays a critical role in chromosomal structure maintenance during meiosis[136], facilitated by the DNA topoisomerase 2-alpha subunit TOP2A, a well-established therapeutic target[137] enabling chromosomal condensation and chromatid separation. FOXM1, CENPF, MYBL2, and TRIP13 were implicated as part of a core "proliferation cluster," associated with poor outcome, whose activity is dependent on p53 inactivation[138]. Indeed, TP53 mutations emerged as the most significant event upstream of MRB:2. Additional proliferation-related proteins, such as E2F2, E2F7, and TIMELESS, contribute to MRB:2's strong association with proliferative hallmarks such as *E2F Targets* ($p = 8.1\times10^{-76}$), *Mitotic Spindle* ($p = 2.6\times10^{-2}$) and *G2/M Checkpoint* ($p = 3.5\times10^{-45}$), as well as *MTORC1* ($p = 1.7\times10^{-5}$) and *V1 and V2 MYC* programs ($p = 1.2\times10^{-28}$ and 3.7×10^{-10}, respectively). Finally, UHRF1, also a candidate therapeutic target, is overexpressed in many cancers[139], where it regulates gene expression and peaks in G1 phase, continuing through G2 and M, while ASF1B—a core member of the histone chaperone proteins, responsible for providing a constant supply of histones at the site of nucleosome assembly and the most recurrent activated MR—is predictive of outcome in several tumors[140]. Thus, while the role of these proteins may have been individually established in some cancers, our study identifies them as a hyper-connected, synergistic core module activated in the most aggressive cancer subtypes, from melanoma and GBM, to colorectal, prostate, and ovarian cancer (**Figure 2.5A**).

Activity of MRB:3 and MRB:7 was also associated with proliferation, yet via complementary MRs such as E2F1/2/7/8 and chromatin remodeling enzymes involved in mitotic

progression (SUV39H1), assembly (CHAF1B), and mini-chromosome maintenance (MCM2/3/6/7).

At the other end of the functional spectrum, MRB:24—significantly associated with *inflammatory response* and *immune related* hallmarks, including via the immune-regulator MR STAT1 (**Figure 2.5B**)—was activated in 20 subtypes (**Figure 2.5A**) and highly predictive of outcome (e.g. in SKCM, **Figure 2.5C**). MRB:19 was also enriched in immune related hallmarks (**Figure 2.5B**) via alternative MRs, including CIITA, an MHC transactivator, whose inactivation abrogates HLA-DR presentation and promotes immune-evasion[141], CD86, the canonical CTLA-4 ligand involved in immune checkpoint activation, and additional proteins (e.g., NOTCH4, MITF, etc.) associated with an immune-evasive microenvironment[114].

Taken together, these data suggest that MRBs may provide complementary "molecular recipes" for implementing the same cancer hallmarks in different tumor contexts.

Obviously, there are several limitations to the MOMA analyses, providing options for potential future improvements. Consistent with other high-throughput methods, both experimental and computational, it is reasonable to expect that MOMA will also produce false positive and negative predictions. Moreover, MOMA was not optimized on an individual cohort basis but rather to identify commonalities across different tumor subtypes. As such, it is not intended as a replacement but rather as a complement to existing analyses, specifically to identify proteins that canalize cancer alterations towards subtype implementation. For instance, TP53 mutations, are ubiquitous in ovarian cancer, thus providing minimal contribution to its subtypes and failing detection by MOMA. Similarly, the proposed clustering strategy may over- or under-stratify some cohorts, in order to avoid missing rare subtypes across most cohorts. For instance, S_6, the most aggressive PRAD subtype (**Figure 2.6A**), would have been missed by a more conservative

clustering strategy. Yet, tuning the algorithm for rare subtypes may cause over-stratification of others. Indeed, while most subtypes are molecularly distinct, PAAD subtypes S_3, S_4, and S_5 were quite similar, both in terms of MRs and upstream genetics. Conversely, under-stratification was evident in breast cancer, where MOMA identified only four subtypes, a basal-like one (S_4), a Luminal-B one (S_2), and two molecularly-distinct Luminal-A ones (S_1 and S_3). Forcing a more granular 8-cluster solution split the basal subtype into Claudin^{low} and Claudin^{high} subtypes (**Figures S2.2D, S2.2E**), HER2 positive tumors, however, still failed to form a separate cluster and were enriched in either the Luminal B or Basal subtypes (**Figure S2.2B**), suggesting that, while HER2+ tumors may present a distinct oncogene dependency, due to their hallmark mutation, their transcriptional identity may be more consistent Basal (HR-negative) and Luminal B (HR-positive) tumors.

Some key events may also be missed (false negatives) due to the highly conservative nature of the DIGGIT analysis. Indeed, BRAF mutations, which are frequent in SKCM, were significantly associated with differential MR activity by aQTL analysis. Yet, they were not identified as upstream MR modulators by CINDy, because activity of this protein is not effectively tracked by VIPER, and were thus missed by MOMA. Indeed, previous validation[43,75] shows that ~20% of proteins harboring functional genetic alteration may be missed by VIPER analysis. We are currently developing approaches to further improve sensitivity, for instance by including DNA binding motifs, ATAC-Seq data, or other epigenetic data modalities. Similarly, as also reported, VIPER may invert the sign of differential activity due to autoregulatory loops. This does not compromise MR identification but may identify some activated MRs as inactivated and vice-versa. Further improvement to the algorithm may be possible by changing the integration logic or by

using mutational or perturbational data to better infer the activity of mutation harboring proteins, as shown in[142].

While the current version of MOMA identified a large repertoire of previously unreported mutations and subtypes, the algorithm may be tuned for improved stratification, on an individual tumor cohort basis, for instance by using the average of each cohort, rather than the average of TCGA, as a control, as shown in several prior studies, e.g., [64,143], thus further highlighting subtle subtype differences.

To make MOMA broadly available to the research community, we deposited the related software in Bioconductor[96], allowing its application to any cohort for which matched gene expression and mutational data is available. We also developed a public-access Web Application that allows biologists to easily query and visualize the ~2 million tumor-specific molecular interactions emerging from the analysis[98].

Acknowledgments

We acknowledge the genomic and small animal imaging shared resources of the Herbert Irving Comprehensive Cancer Center, supported in part by P30CA013696. This work was also supported by the National Cancer Institute's (NCI) Office of Cancer Target Discovery and Development (CTD^2) initiative (U01CA217858) to AC, a NCI Outstanding Investigator Award (R35CA197745) to AC, the NCI Research Centers for Cancer Systems Biology Consortium (U54CA209997) to AC, the Prostate Cancer Foundation grant 18CHAL07 to AC, NIH R01 (R01CA173481 and R01CA196662) to CAS, and the NIH Instrumentation grants (S10OD012351 and S10 OD021764) to AC. A.A. was supported by grants from the Spanish ISCIII-MINECO (PI19/00342; PI16/01070), EAURF/407003/XH, a DOD Award (W81XWH-

18-1-0193) and the CERCA Program/Generalitat de Catalunya, and FEDER/ERDF funds - a way to Build Europe. A.V. is supported by the DOD Early Investigator Research Award (W81XWH19-1-0337).

Author Contributions

Conceptualization and Methodology, E.O.P., A.A., F.M.G., M.J.A., C.AS. and A.C.; Investigation and Formal Analysis, E.O.P, A.A., P.S., F.M.G, E.F.D., S.J.J., A.V., M.J.A., and A.C.; Resources and Software, E.O.P., B.C., S.J.J., S.T., and P.S.; Writing – Original Draft, E.O.P, P.S., and A.C.; Writing – Review and Editing, all authors.

Declaration of Interests

A.C. is founder, equity holder, consultant, and director of DarwinHealth Inc., a company that has licensed some of the algorithms used in this manuscript from Columbia University. M.J.A. is Chief Scientific Officer and equity holder at DarwinHealth, Inc. Patent 10,790,040, titled "Virtual Inference of Protein Activity by Regulon Analysis" has issued on Sept. 29, 2020 related to the VIPER method. Columbia University is also an equity holder in DarwinHealth Inc.

2.5 Methods

2.5.1 Key Resources Table

Table 2.1: Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit Anti-GRHL2	Millipore Sigma	Cat# HPA004820
Rabbit Anti-SPDEF	Proteintech	Cat# 11467-1-AP
Rabbit-Anti AR	Cell Signaling Tech.	Cat# 5153S
Mouse Anti-□-catenin (JUP)	BD Biosciences	Cat# 610253
Mouse Anti-E-Cadherin (CDH1)	BD Biosciences	Cat# 610404

Chemicals, Peptides, and Recombinant Proteins		
Fedratinib	Selleck Chemicals	Cat# S2736
Pevonedistat	Selleck Chemicals	Cat# S7109
Lexibulin	Selleck Chemicals	Cat# S2195
ENMD-2076	Selleck Chemicals	Cat# S1181
Triapine	Selleck Chemicals	Cat# S7470
Dorsomorphin	Selleck Chemicals	Cat# S7306
Raltitrexed	Selleck Chemicals	Cat# S1192
Deposited Data		
TCGA Sample Data	Broad Institute	https://gdac.broadinstitute.org/
PRADA Gene Fusion Data	The Jackson Laboratory	https://www.tumorfusions.org/
Achilles shRNA Essentiality Data	DepMap; Broad Institute	https://depmap.org/portal/achilles/
METABRIC Breast Cancer Patient Data	cBioPortal; Curtis et al., 2012	https://www.cbioportal.org/study/summary?id=brca_metabric
Pancancer Driver Genes	Bailey et al., 2018	https://doi.org/10.1016/j.cell.2018.02.060
Network of Cancer Genes (NCG)	Repana et al., 2019	http://ncg.kcl.ac.uk/
Molecular Signatures Database (MSigDB)	UC San Diego; Broad Institute	https://www.gsea-msigdb.org/gsea/msigdb/index.jsp
Gene Ontology	Gene Ontology Consortium	http://geneontology.org
Enzalutamide-treated LNCaP cells	Handle et al., 2019	GEO Accession# GSE130534
Analysis of prostate cells and tumor biopsies	Rajan et al., 2014	GEO Accession# GSE48403
Analysis of prostate cells and tumor biopsies	Zhang et al., 2016	GEO Accession# GSE067070
Experimental Models: Cell Lines		
LNCap clone FGC	ATCC	Cat # ATCC® CRL-1740
DU 145	ATCC	Cat# ATCC® HTB-81
22Rv1	ATCC	Cat# ATCC® CRL-2505
PC-3	ATCC	Cat# ATCC® CRL-1435
293 [HEK-293]	ATCC	Cat# ATCC® CRL-1573
Experimental Models: Organisms/Strains		
Immunodeficient Athymic Nude mice - Foxn1^{nu}	Envigo	Model# Hsd:Athymic Nude-Foxn1^{nu} --069
Oligonucleotides: shRNA Clones		
See Table S5 for clones		
Recombinant DNA		
pMD2.G	Laboratory of Didier Trono via Addgene	Addgene plasmid #12259
psPAX2	Laboratory of Didier Trono via Addgene	Addgene plasmid # 12260
Software and Algorithms		
MOMA Web application	This paper	http://www.mr-graph.org/
MOMA Bioconductor Package	This paper	https://bioconductor.org/packages/release/bioc/html/MOMA.html
R for Statistical Programming	R Core Team, 2020	https://www.R-project.org/
Complex Heatmap	Gu et al., 2016	https://doi.org/10.1093/bioinformatics/btw313
Q-Value Estimation for FDR	Storey et al., 2020	http://github.com/jdstorey/qvalue
ggplot2: Graphics for Data Analysis	Wickham et al., 2016	https://ggplot2.tidyverse.org

VIPER R package	Alvarez et al., 2016	https://doi.org/10.18129/B9.bioc.viper
mixtools R package	Benaglia et al., 2009	https://www.jstatsoft.org/article/view/v032i06
DEBrowser	Kucukural et al., 2019	https://debrowser.umassmed.edu/
clusterProfiler R package	Yu et al., 2012	http://yulab-smu.top/clusterProfiler-book/
MutSig2CV	Lawrence et al., 2013	https://software.broadinstitute.org/cancer/cga/mutsig
Mutation Assessor	Reva et al., 2011	http://mutationassessor.org/r3/
CHASM	Carter et al., 2009	https://wiki.chasmsoftware.org
GISTIC 2.0	Mermel et al., 2011	https://doi.org/10.1186/gb-2011-12-4-r41
PrePPI	Zhang et al., 2012	https://honiglab.c2b2.columbia.edu/PrePPI/index.html
HumanNet v2	Hwang et al., 2018	https://www.inetbio.org/humannet/
Multinet	Khurana et al., 2013	https://doi.org/10.1371/journal.pcbi.1002886

2.5.2 Resource Availability

Primary Dataset Information

Source data for the analyses done in the paper is available from the TCGA Firehose Repository (gdac.broadinstitute.org, 2016-01-28 release). Full description of data types per sample (RNA sequencing, SNV and SCNA) acquired from TCGA firehose available in Supplemental Table 1. All samples with RNA sequencing data available were used in the analysis. Cohorts with fewer than 100 samples were not used. Further information about sample acquisition and relevant clinical annotations are available on the TCGA website. Fusion data was acquired from the Tumor Fusions Gene Data Portal, which is based on the TCGA data (www.tumorfusions.org, 2017-10-01 release)[99,144].

Results Data

The results of the analysis can be interactively accessed on our MOMA web application (http://www.mr-graph.org/). Code used to analyze the data has been compiled into a Bioconductor R package, MOMA, that can be downloaded here (https://bioconductor.org/packages/release/bioc/html/MOMA.html).

2.5.3 Experimental Model and Subject Details

Animals

The immunodeficient NCr nude Spontaneous mutant model (Envigo; Product model: Mutant mice - Hsd:Athymic Nude-Foxn1^{nu} - 069) was used for the MRB:2 xenograft validation experiments. All experimental procedures were approved by the Ethical Committee on Animal Research at IDIBELL, and have been authorized by the responsible Department of the Catalan Autonomous Government (File Number: FUE-2016-00307059; Project Number: 9025, Project coordinator: Alvaro Aytés). The barrier facility at IDIBELL is an AAALAC-certified facility. Maximum cage density was 5 mice/cage and cages were placed in ventilated racks with water ad libitum and chow replenished weekly as well as clean new bedding. All animals used in this study were 6 weeks old male athymic Nude-Foxn1nu (Envigo). Mice were monitored daily for signs of distress throughout the course of the experiment.

Cell lines

All cell lines were acquired from ATCC, as authenticated by them. Growth medium for cells is as follows: LNCaP cells and 22Rv1 cells were grown in RPMI-1640 medium (Gibco) supplemented with 10% Fetal Bovine Serum (FBS; Sigma-Aldrich) and antibiotics (penicillin/streptomycin, P/S;

= 100 units of penicillin and 100 μg of streptomycin per ml of medium); DU145 cells were grown in Eagle's Minimal Essential Medium (Gibco) supplemented with 10 % FBS and P/S; PC3 cells were grown in Ham's F-12K (Kaighn's) Medium (Gibco) supplemented with 10 % FBS and P/S; HEK-293 were grown in DMEM supplemented with 10 % FBS and P/S. All cell lines were grown at 5% CO2 and 37C.

2.5.4 Methods

Sequencing Data and Activity inference

RNA-Seq raw gene counts were downloaded from the TCGA firehose web site (gdac.broadinstitute.org, 2016-01-28 release), transformed to Reads Per Kilobase of transcript, per Million mapped reads (RPKM), using the average transcript length for each gene and $\log_2$ transformed. Transcriptome-wide expression signatures were computed by two non-parametric transformations. First, each column (tumor sample) was rank transformed and scaled between 0 and 1. Then each row (gene) was rank transformed and scaled between 0 and 1. Finally, regulatory protein activity was measured by the VIPER algorithm[75], using tissue-matched ARACNE regulons[145,146] (See **Figure S2.1B**).

Systematic experimental validation has confirmed that VIPER can accurately measure differential activity for >80% of transcriptional regulator proteins, when ≥ 40% of the genes in a regulon represent bona fide targets of the protein[75]. In addition, multiple studies have experimentally validated that >70% of ARACNe-inferred targets represent bona fide, physical transcriptional targets—e.g., by Chromatin Immunoprecipitation (ChIP) and RNAi-mediated silencing, followed by gene expression profiling[63–65,75]—thus fulfilling the VIPER requirements for accurate protein measurement. The results of the VIPER analysis are reported as a Normalized

Enrichment Scores (NES) values of a protein targets in differentially expressed genes with respect to the centroid of TCGA, as assessed by aREA (see below). This has been shown to accurately characterize differential protein activity. Positive NES values (shown as a red gradient) indicate increased protein activity while negative NES values (shown as a blue gradient) indicate decreased protein activity.

Genomic events

Candidate genomic event data were downloaded from the TCGA firehose (gdac.broadinstitute.org, 2016-01-28 release). For mutations and small indels, we downloaded Mutation Annotation Files (MAF) and selected all events annotated as non-silent alterations. For SCNAs, we downloaded SNP6 copy number profiles and selected a threshold of +/-0.5 as the value that provides an optimal tradeoff between sensitivity and specificity in capturing copy number changes, as discussed in the literature[147].

To ensure that copy number changes are functionally relevant, we adopted the approach discussed in the DIGGIT manuscript[83]. Specifically, only SCNA genes whose correlation between copy number and expression was statistically significant across a cohort were considered as functional candidates (**Figure S2.1B**). For the Genomic Saturation analysis, GISTIC2.0 results were downloaded from Firehose to better account for proximal copy number alteration events and to differentiate between focal (score of +/-2) and regional (score of +/-1) events. When multiple functional events were identified within the same amplicon, they were consolidated into a single event vector, thus preventing overcounting (Region Consolidation). However, for completeness, the MOMA Web App reports the identity of all events in an amplicon that pass the CINDy and

aQTL analyses. Finally, gene-fusion calls were called by the PRADA algorithm, and downloaded from the Tumor Fusions Gene Data Portal (www.tumorfusions.org, 2017-10-01 release)[99,144].

aREA Analysis

The analytic Rank-based Enrichment Analysis (aREA) was introduced in[75] as an analytical methodology to assess gene set enrichment analysis statistics, producing results that are virtually identical to GSEA[50] without the need for time-consuming sample or gene shuffling.

DIGGIT Analysis

We implemented a slightly improved version of the DIGGIT algorithm. The original DIGGIT combined (a) a MINDy analysis step[82] to identify proteins representing candidate upstream modulators of a MR protein (b) an aQTL analysis step to identify genomic events in candidate upstream modulators associated with statistically significant differential MR activity, and (c) a conditional association analysis step to eliminate genomic events that were no longer significant given another genomic event. The analysis was improved as follows: (a) rather than using mutual information, aQTL statistical significance is assessed by aREA-based enrichment analysis of samples, ranked by differential activity of the specific MR, in samples harboring a specific SNV or SCNA events, (b) the MINDy algorithm was replaced by CINDy[79], providing a more accurate implementation of the conditional mutual information foundation of the algorithm, and (c) the conditional association analysis step was eliminated because it produced too many statistical ties when applied to pancancer cohorts; note that aQTL analysis was performed only for events occurring in ≥ 4 samples since fewer events are highly unlikely to achieve statistical significance (**Figure S2.1C Step 2**). The individual steps are described in the following.

CINDy Score

Step 1: Proteins were first ranked by their VIPER statistical significance, integrated across all cohort samples using the Stouffer's method for p-value integration[148].

Step 2 For each statistically significant differentially active protein (i.e. candidate MR) the *conditional mutual information* $CMI = I[MR, \{T_i\}|M]$, between the expression of the MR and of its regulon genes, given the expression of any gene harboring a somatic event, was computed. Thus, CINDy identified mutation-harboring genes encoding for proteins that affect the ability of a MR to regulate its targets (**Figure S2.1B**).

Step 3: For each event type (i.e. SNV, amplified SCNA, or deleted SCNA) all statistically significant CINDy scores for a given MR were integrated using Stouffer's method to produce three global CINDy scores $S_C^{\mathrm{SNV}} = -\mathrm{Log}_{10}(p_C^{\mathrm{SNV}})$, $S_C^{\mathrm{Amp}} = -\mathrm{Log}_{10}(p_C^{\mathrm{Amp}})$, and $S_C^{\mathrm{Del}} = -\mathrm{Log}_{10}(p_C^{\mathrm{Del}})$. Fusion events were not analyzed in this fashion since ARACNe is not designed to identify targets of fusion proteins. Thus, for fusion events, only the aQTL analysis step was applied.

aQTL Score

Step 1: Proteins were ranked by their VIPER statistical significance, integrated across all cohort samples using Stouffer's method. This could be further improved in the future by integrating across individual subtypes rather than entire cohorts.

Step 2: For each statistically significant differentially active protein (i.e. candidate MR) and somatic event (SNV, SCNA, or FUS), the statistical significance of the aQTL event was assessed by computing the enrichment of all cohort samples, ranked by the MR's differential activity, in samples harboring the event, using aREA.

Step 3: For each event type, a global aQTL score (S_{aQTL}) was computed as the $-Log_{10}(P_{aQTL})$, with P_{aQTL} representing the integration of all statistically significant MR-event aQTL p-values ($p \leq 0.05$) per MR for that event type, using Stouffer's method. This produced three global aQTL scores S_{aQTL}^{SNV}, for SNVs, small indels, and fusion events, S_{aQTL}^{Del}, for SCNA deletion, and S_{aQTL}^{Amp} for SCNA amplifications. If ≥ 100 CINDy-inferred MR modulators were identified in a given cohort (see CINDy Score), then only aQTLs for somatic events harbored by genes with a statistically significant CINDy p-value were integrated. Otherwise, the p-values of all statistically significant aQTLs were integrated independent of CINDy results. This is because fewer than 100 statistically significant CINDy modulators indicates that the dataset is too small for a properly powered CINDy analysis.

PrePPI Score

PrePPI[84] is used to identify structure-based protein-protein interactions between proteins encoded by genes harboring a somatic event and each MR protein.

Step 1: Proteins were first ranked by their VIPER statistical significance, integrated across all cohort samples using Stouffer's method.

Step 2: High-confidence interactions in the PrePPI database 1.2.0 (likelihood > 0.5) were assigned an empirical *p*-value as follows: first they are ranked based on their likelihood scores; then *p*-values were computed as the fraction of interactions with equal or better rank, normalized by the total number of PrePPI interactions in the database.

Step 3: For each event type, a global PrePPI score (S_P) was computed as the $-Log_{10}(P_{PrePPI})$, with P_{PrePPI} generated by integrating the individual *p*-values of all statistically significant PrePPI interactions ($p \leq 0.05$) for that event type, using Fisher's method (Jerby-Arnon et al., 2014). This produced three global PrePPI scores S_{PrePPI}^{SNV}, S_{PrePPI}^{Del}, and S_{PrePPI}^{Amp}.

Integrated rankings and MOMA Scores

Step 1: For each candidate MR, the *p*-values corresponding to same-type events (e.g., all SCNA deletions) as assessed by aQTL, PrePPI, and CINDy, were integrated using Stouffer's method. For fusion events, CINDy and PrePPI scores cannot be computed and are thus not integrated. For the aQTL analysis, fusion events were considered equivalent to SNVs. This produced 9 integrated *p*-values for each statistically significant, candidate MR protein: p_{aQTL}^{SNV}, p_{aQTL}^{Amp}, p_{aQTL}^{Del}, p_{PrePPI}^{SNV}, p_{PrePPI}^{Amp}, p_{PrePPI}^{Del}, p_{CINDy}^{SNV}, p_{CINDy}^{Amp}, and p_{CINDy}^{Del}.

Step 2: After ranking all proteins in a cohort based on their VIPER score, we used Stouffer's method to integrate the 9 *p*-values for each statistically significant protein (i.e., candidate MR) with its VIPER *p*-value, thus creating a global MOMA *p*-value ($p_M(MR)$). The latter representing the probability that a protein may be a *bona fide* MR by chance. A global MOMA score was then computed as $S_M(MR) = -\text{Log}_{10}\big(p_M(MR)\big)$ squared (**Figure S2.1C**).

Cluster Reliability Score (CRS)

The CRS was introduced in[130] as a statistically sound way to assess the fit of each sample within a cluster. For each sample, a distance vector $\mathbf{V}_1$, representing its distance from all other samples in the same cluster and a vector $\mathbf{V}_2$, representing its distance from all other samples in the cohort are computed. The sample distance matrix was computed by taking the weighted VIPER scores for each sample (VIPER activity values multiplied by each MR's MOMA Score) and calculating the pairwise Pearson correlations. The normalized enrichment score of $\mathbf{V}_2$ distances, ranked from the largest to the smallest one, in $\mathbf{V}_1$ distances, is then assessed using aREA. This produces a *p*-value that represents the tightness and separation of the cluster being considered in relation to all other samples. A *cluster-wide reliability score* for each cluster is assessed as the average cluster reliability (NES) of each sample in the cluster, scaled between 0 and 1. Finally, the reliability of the entire clustering solution (*global cluster reliability score*) is assessed as the average of the cluster-wide reliability score of all clusters in the solution.

Activity-based Clustering

Each tissue-specific VIPER activity matrix was clustered using *k*-medoids clustering, with *k* ranging from 2 to 10 clusters, using a distance matrix defined by the weighted Pearson correlation between VIPER-inferred protein activity vectors. Weights were defined as the square of the integrated MOMA scores ($S_M^2(MR_i)$), thus increasing the contribution of high-scoring MRs (**Figure S2.1D**). Cluster Reliability Scores (CRS) were calculated for each sample and for each *k* value and the optimal number of clusters was determined as the first local maximum for the Global Cluster Reliability Score. We used a Kolmogorov–Smirnov test between the CRS of the samples from the optimal *k*-cluster solution (i.e. the one with the highest global reliability score) and the

CRS of the samples from every other *k*-cluster solution to identify solutions that were statistically indistinguishable. Among those, we selected the one producing the best survival separation, as described in *Survival analysis*.

Silhouette Scores

Silhouette Scores were computed as described in[102]. They were used purely for visualization purposes, since they are well-established as metrics to assess cluster reliability.

Expression-based Clustering

Similar to Protein Activity-based clustering, each tissue-specific gene expression matrix was clustered using *k*-medoids clustering with *k* set as the same value chosen for the tissue-specific VIPER activity clustering. Distance between samples was defined using Pearson correlation between gene expression profiles. Cluster Reliability Scores and Silhouette scores were computed as described in above.

Survival analysis

Clinical data was downloaded from the Broad Institute GDAC website (gdac.broadinstitute.org). We used the 'survival' R/CRAN package version 2.41-3 to fit a Cox proportional hazards model to each sample grouping defined by the initial clustering. We then defined the "best" survival clusters as the one with the lowest proportion of observed to expected death events, and the "worst" survival as the highest observed/expected ratio. We then fit a second Cox model exclusively to samples from those two clusters and calculated the significance of survival differences between "best" and "worst" clusters in that model.

Saturation Analysis

Saturation curves were generated by ascertaining the number of functional somatic events upstream of the *N* most statistically significant candidate MR proteins, ranked by their global MOMA score. To assess an appropriate saturation threshold, we first assessed how many functional somatic events $N_{E=1,253}$ were upstream of the first half (1,253) of all regulatory proteins in that subtype, thus conservatively excluding proteins with a non-statistically significant VIPER activity. The saturation threshold then was set at 85% of that number $N_0 = 0.85 \times N_{E=1,253}$. We then assessed how many of the *N* proteins with the highest VIPER activity were needed to identify N_0 somatic events in their upstream pathways. For all subtypes—except for 3 Ovarian cancer subtypes (S_1, S_3 and S_4)—saturation increased so rapidly and significantly, compared to an identical number of randomly selected regulatory proteins (null hypothesis), that increases in event number for $N > 100$ MRs were not statistically significant. To avoid contaminating functional genomic events with passenger ones, by using non-significant MRs to assess saturation, we thus selected a more conservative saturation threshold $N_1 = 0.85 \times N_{E=100}$. We used N_1 for all subtypes except for the three ovarian cancer subtypes for which we used N_0.

Genomic Plots

To visually represent genomic events upstream of MR proteins in each sample, as identified by saturation analysis, we used cBioPortal OncoPrint[149], with ComplexHeatmap[109]. To avoid clutter, we restricted visualization to events previously reported as oncogenes and tumor suppressors[33,150]. However, all events can be downloaded from the MOMA Web App. For amplified or deleted SCNAs, we determined whether an oncogene or tumor suppressor had been identified by MOMA

as functional in that region, before region consolidation (see *Genomic Events*). For regions with a single oncogene/tumor-suppressor its name is used as representative of the SCNA. When two were detected, their names separated by a semicolon were used. When three or more were detected, the SCNA locus is used followed by "-multi." Due to size constraints for figure representation, a maximum of 50 most frequent events is shown. However, complete driver event lists are available on the MOMA Web App. The option to generate OncoPrint plots with all genes is prioritized for the next version of the application.

Driver Mutation Enrichment

To assess the statistical significance of somatic event enrichment, upstream of checkpoint MRs, we performed a sample-specific analysis in each cohort. For each sample we identified activated MRs and their upstream somatic events using the same methodology described in the *Saturation Analysis* section. Then, for each sample, we computed the ratio of all validated CHASM[151] and GISTIC2.0[106] putative driver events vs. the total number of events (**Figure S2.3C**). To assess the cohort-level significance, we compared the number of samples with a ratio > 1 against a one-tailed binomial null distribution ($p = 0.5$). This showed that every cohort but one (LAML) showed significant enrichment in putative driver genes (**Figure S2.3B**).

MRB Analysis

The 407 MRs identified by saturation analysis that were also statistically significant in ≥ 4 subtypes (recurrence analysis) were clustered based on their VIPER-inferred activity, using a Euclidean distance metric and partitioning around medoids (PAM) for $k = 2$ to 100 clusters (**Figure S2.1E**). To compute the Euclidean distance, each MR was associated with a 112-

dimensional vector representing its VIPER-inferred activity in each subtype. A Cluster Fitness score was defined as the Average Cluster Reliability Score for all MRs in a cluster. The analysis identified k = 24 as the optimal clustering solution (**Figure S2.5A**). Each "core-set" cluster identified by this analysis was then expanded by the *m MRs* with the best average Euclidean distance to those in the core-set, for $m = 0, \ldots 100$. For each *m* additional MRs in each MRB, the trace of the covariance matrix of the Tumor Hallmark enrichment across the 24 MRBs was calculated to assess the total variance of the solution. This variance showed optimal increase for $m = 6$ (**Figure S2.5B**). These optimization steps to ensured uniqueness, specificity, and robustness of the MRB solution.

Jaccard concordance index

Each MRB is represented as a 112-dimensional vector representing its statistically significant activation (1), inactivation (-1) or neutral (0). The Jaccard concordance index between two MRBs is the scalar product of their associated vectors, such that co-activation or co-inactivation of the MRB in the same subtype increases the score by 1 while non-concordant activity in a subtype does not increase the score.

MRB Enrichment Analysis

Cancer Hallmarks include 50 gene-sets defined by the Broad Institute and refined/simplified by others[49,111]. To calculate downstream enrichment, we pooled genes from the regulons of each MR in each MR block that had a highly significantly likelihood of being a physical target ($p < 0.05$) and that were identified in at least 2 different tissues. We then assessed enrichment using the hypergeometric distribution between MR targets and each Hallmark's gene set. The same

approach was used to compute enrichment in KEGG and Reactome gene sets. Significance was assessed by Benjamini-Hochberg False Discovery Rate (FDR) to account for multiple hypothesis testing. Only significant enrichments (FDR < 0.05) are shown. To calculate enrichment of genomic events upstream of MR blocks, we selected the top 100 most significant predicted upstream genomic events, for both SNVs and functional SCNA genes, in subtypes with significant MRB activity ($p < 10^{-3}$). The hypergeometric overlap between these gene sets and the Hallmark, KEGG and Reactome gene sets was performed as described above. A fixed event number was chosen to avoid biasing the statistical analysis for MRBs with a greater number of upstream events. All enrichment analyses were done using the enricher function from the R clusterProfiler package[152].

Achilles Essentiality

Achilles shRNA DEMETER knockout scores were downloaded from The Broad Institute for all cell lines in CCLE for all TFs and co-TFs analyzed by MOMA. To identify a natural threshold to assess essentiality, Achilles dependency scores were re-normalized by fitting a bimodal normal mixture models using the R package 'mixtools' [153]. The normal probability density with the most positive (i.e., least essential) mean was set as the null-hypothesis (*essentiality null hypothesis probability density*) to assess essentiality as a z-score. This allows setting an appropriate null hypothesis to assess essentiality on a gene by gene basis.

For each of the 112 MOMA subtypes, we matched the MR activity vector, weighted by the cohort-specific MOMA score of each MR, to the protein activity profile of each CCLE cell line, using the 'viperSimilarity' algorithm included in the VIPER algorithm[75], thus identifying the cell lines that best recapitulates subtype-specific MRs as possible dependencies. We then assessed the essentiality of each MR in cell lines that were significant matches ($p < 0.01$; Bonferroni correction)

vs. those providing clear non-matches ($p = 1$) using a non-parametric rank-based Mann-Whitney-Wilcox test based on the null hypothesis probability density defined in the previous paragraph; significant FDRs after multiple hypothesis correction (Benjamini-Hochberg FDR < 0.05) were considered essential subtype-specific MRs. Essentiality was then stratified for each MR across the subtypes where that MR was statistically significantly active. To calculate statistical significance of the enrichment of essential genes, a null model was built by taking 10^6 random selections of MRs equivalent to the number of MRs in each tumor checkpoint and then counting the number of essential MRs across all subtypes. These permutations were then fitted to a normal distribution (**Figure S2.5F**).

METABRIC Breast cancer analysis

ARACNE was run with 100 bootstrap iterations and a mutual information significance threshold of $p = 10^{-8}$, separately for candidate TF and coTF regulators, using METABRIC gene expression profile data. For each sample, protein activity was inferred using VIPER. Survival analysis was performed by first calculating the mean VIPER activity across checkpoint proteins and binning samples into "high" and "low" quantiles, for each checkpoint. Clinical data was downloaded from the cBioPortal. We used the 'survival' R/CRAN package version 2.41-3 to fit a Cox proportional hazards model to each sample grouping, using the last known follow-up date, and testing for significant survival differences with that model.

MRB:2 Analysis

For each of the candidate Master Regulator proteins in MRB:2 we computed the rankings based on the integrated *p*-value of each MR-event in prostate cancer, as well as the cross-pancancer

rankings for the same interactions. For each MR and each somatic event, *p*-values were generated as discussed in the DIGGIT methods section. A joint rank from these two lists was then created using an additive mean and the top 20 interactions were retained for each MR. These Interactions were visualized as a network graph (**Figure 2.6F**) with the Cytoscape software package[154]. Network edges between MRB:2 proteins and mutation events identified in **Figure 2.6F** were included in the sample/event plot (**Figure 2.6E**). Events with significant copy number associations were also included if they contained one or more samples with a mutation in that same protein. Additionally, for interactions with only copy number events (deletion, amplification) we computed the aREA association score with the average activity of MRB:2, and selected the top 10 most significant deleted and amplified genes, respectively, to include on the plots.

MRB:2 Validation

Lentiviral-mediated gene silencing

Silencing of SORBS3, BCAR1, MAP3K7, PTEN, Tp53 was achieved by lentiviral delivery of validated shRNAs. Two target-specific shRNAs in the pLKO.1 lentiviral vector were co-transfected in HEK-293 cells together with the pMD2.G and psPAX2 envelope and packaging plasmids in 1% FBS. pMD2.G and psPAX2 were gifts from the laboratory of Didier Trono (Addgene plasmid # 12259; http://n2t.net/addgene:12259; RRID:Addgene_12259 and Addgene plasmid # 12260; http://n2t.net/addgene:12260; RRID:Addgene_12260) Supernatants were recovered at 24 and 48 hours and were later concentrated using the Lenti-X concentrator reagent (Takara #631231). The 22Rv1 human prostate cancer cell line was spin-infected at multiplicities of infection (MOI) of approximately 1 in the presence of 8 μg/mL polybrene (hexadimethrine bromide), then incubated with virus for approximately 18 hours in a 37°C, 5% CO2 incubator. At

48h post-infection, cells were selected with 2 µg/mL puromycin and at 96h post-transduction medium was changed to fresh complete medium. Efficiency of gene silencing was assessed by qPCR using primers for each of the targets and comparing target expression against cells transduced with the MISSION® Non-Target shRNA Control Transduction Particles.

Perturbation dataset VIPER analysis

To assess the effect of selected gene silencing on MRB:2 MRs, we generated a signature for count data from each experimental condition, using the control condition as a reference, and performing a t test, using 100 permutations of the samples (columns) as a null model. This signature and null model were inputted to the 'msviper' function in the VIPER Bioconductor package, along with the TCGA Prostate cancer regulon. A second null model was constructed by re-running this same analysis on 100 permutations of the column labels, and a t-test was performed between the VIPER scores from each condition and this null, to assess the overall ability in reverting the signature for checkpoint 2 proteins.

Wound Healing Assays

Control and silenced cells where seeded at high concentration in 6 well plates in triplicate using a silicone insert. At day 1 the silicone insert was removed and cell migration into the gap was monitored at 24h, 48h and 72h hours. The percent of migrating cells was quantified, relative to non-targeting controls, by measuring the cell-free area with ImageJ software. A Mann–Whitney U test was used to calculate the significance (P value) of the difference between the control (n=3 replicates) and knockdown cells (n= 6 replicates; 3 for shRNA shRNA#1 and 3 for shRNA#2)

Matrigel invasion assays

5×10^4 cells were seeded in the BD FluoroBlok inserts (BD Biosciences) in FBS-free media. Inserts were placed in 24-well plates containing RPMI supplemented with 10% FBS as chemoattractant. Invasion was monitored using a bottom-reading fluorescence plate reader and invading cells detected using calcein AM fluorescent labeling. The fluorescence signal was quantified with ImageJ, and a Mann–Whitney U test was used to calculate the significance (*p*-value) of the difference between the control (n=3 replicates) and gene-silenced cells (n= 6 replicates; 3 for shRNA shRNA#1 and 3 for shRNA#2).

Xenograft assays

IDIBELL´s Institutional Animal Care and Use Committee (IACUC) had approved all animal procedures. For analyses in vivo, 5×10^6 22Rv1 cells expressing the control or target shRNA lentivirus were mixed with Matrigel (1:1 vol/vol) and injected into the right flank of immunodeficient nude mice (Envigo, Nude-Foxn1^{nu}); tumor growth was monitored with calipers until one of the experimental groups reached the maximum 1.5 cm^3 tumor volume. One-way analysis of variance (ANOVA) was used to calculate statistical significance (*p*-value) of the difference between control and silenced groups.

MRB:14 Validation

Analysis of Enzalutamide-treated LNCaP cells

Gene counts for this dataset were downloaded from Gene Expression Ominibus, (GEO), accession GSE130534[123]. Analysis of the counts were performed using the DEbrowser tool[155].

MRB:14 Drug Prioritization

We used a dataset of protein activity profiles of drug response, as inferred from a screening of 120 FDA-approved drugs and 217 late-stage experimental compounds (in Phase 2 and 3 trials) in the DU145 prostate cancer cell line. Profiles were generated at 24h following perturbation with the compound's IC20 concentration determined at 48h by 7-point dose response curves. This concentration was selected to represent the highest sub-lethal concentration that would help elucidate the compound mechanism of action without significantly triggering additional cell response mechanisms, e.g., associated with drug stress response or cell death, that would confound the analysis.

The aREA function from the R VIPER package 1.20.0 was used to compute a Normalized Enrichment Score (NES) for each drug, based on the enrichment of differentially activated proteins, as inferred by VIPER, in MRB:14 MRs. NES values were converted to *p*-values and corrected for multiple hypothesis testing, using the Bonferroni method. Finally, $-\text{Log}_{10}\, p$ was used as a score to prioritize drugs and statistically significant drugs, with scores greater than two, were considered as potential candidates to elicit MRB:14 activation.

Analysis of prostate cells and tumor biopsies

Gene expression data (counts) from two studies[124,125] were collected. Both studies were analyzed in the same way as follows. Counts downloaded from the GEO portal (GEO accession GSE48403, and GSE067070) were normalized using the variance stabilizing transformation function available from the DESeq2 package 1.26.0 in R. The metaVIPER approach[156], available from the R VIPER package, was then used to generate two interactomes from the TCGA PRAD cohort (this manuscript) and the 2015 SU2C metastatic Castration Resistant Prostate Cancer (mCRPC)

cohort[157]. Regulons were pruned to the top 100 targets with the highest likelihood using the pruneRegulon function of the VIPER package. Gene expression signatures for each individual sample were computed using the method *ttest* available from the viper function. Enrichment analysis on VIPER-inferred protein activity signatures was computed and resultant NES scores used. Clustering of labeled samples due to similar activation profiles of MRB:14 on patient samples was performed using the hierarchical clustering algorithm available from the ComplexHeatmap package.

BRCA and BLCA enrichment in MRB:14 activity

Data for PAM50 annotation and luminal/basal subtyping from two studies on TCGA BRCA[158] and BLCA[159] were downloaded. Protein activity profiles for the TCGA BRCA and BLCA cohorts were computed and enrichment scores for MRB:2 and MRB:14 derived. MRB:2, which is a proliferation-associated block (described above), was used as a control. Patients were sorted based on activity NES scores to show correlation between high MRB:14 activity and luminal subtypes as determined by published PAM50 classifiers.

Additional reagents

Small molecule compounds were purchased from Selleck Chemicals (Houston, TX). Culture inserts for migration studies were from iBidi (Gräfelfing, Germany, #80209).

Western Blotting

Cell pellets were lysed in buffer composed as follows: 50mM Tris-HCl, pH 7.5; 250 mM NaCl; 50 mM NaF; 10 mM Na-pyrophosphate; 2.5mM EDTA; 2.5 mM EGTA; 2 mM sodium

orthovanadate; 2% CHAPS; 0.5% Triton-X100; Phosphatase cocktail 3 from Sigma at 1:15 dilution; Protease cocktail (Pierce) 1:15 dilution. After SDS-PAGE separation of equal amounts (~30 ug) of protein lysate from each sample, proteins were transferred to PVDF membranes and then probed with antibodies using standard procedures. Primary antibodies were as follows: AR (Cell Signaling Technology, # 5153S); GRHL2 (Millipore Sigma, #HPA004820); SPDEF (Proteintech, #11467-1-AP); γ-catenin/JUP (BD Biosciences, #610253); CDH1 (BD Biosciences, #610404), diluted 1:1000 each.

Wound Healing Assays

These assays were performed using manufactured cell culture inserts with a defined cell-free gap (iBidi) in 6-well plates. DU145 cells were plated in the inserts at 4×10^5 cells per ml (70 uL per channel). At 24 hrs after plating cells images of the gap were taken (T = 0) and medium was replaced with medium containing the drugs, or DMSO as vehicle control. All drugs were tested at their EC_{50} concentration, 7.2 μM, 1.2 μM, 44 nM, and 8.77 μM for fedratinib, pevonedistat, lexibulin, and ENMD-2076, respectively. Negative control drugs were also tested at their EC_{50} concentration, 1.65 μM, 3.5 μM, and 28nM for triapine, dorsomorphin, and raltitrexed, respectively. After 24 hrs (T = 24), additional images ($n \geq 3$) were taken along the full length of the gap for each treatment. Images were analyzed using the MRI Wound Healing Tool macro (http://dev.mri.cnrs.fr/projects/imagej-macros/wiki/Wound_Healing_Tool) installed in ImageJ. Total gap area was calculated per image and averaged across images for a given sample and converted to % gap remaining (see **Figure 2.7H, 2.7I** legends).

Chapter 3: Multi-omic Analyses of Gastroesophageal Cancer

3.1 Introduction

3.1.1 Gastroesophageal Incidence and Treatment

Globally an estimated 1.5 million new cases of gastric and esophageal cancer have been diagnosed in 2018, making them collectively one of the most commonly diagnosed malignancies in the world[3,160]. As of now, available treatment options are minimal and recurrence is very common, thus contributing to dismal outcome for these tumors, with close to 1.3 million deaths in 2018[3,160]. In particular, targeted therapy options are limited and immune checkpoint inhibitors have so far proven only partially effective. Until the mid 1990s, gastric adenocarcinoma was the leading cause of cancer-related death in the world. However, incidence of this disease has declined over the past several decades, primarily due to decreases in *H. pylori* infection prevalence in more developed countries as well as improvements in food preservation and preparation and early screening in Asian countries[161]. Obesity, age, alcohol consumption, acid-reflux disease and smoking have been associated with increased risk of both gastric and esophageal cancer incidence. As a result, aging populations, with increased prevalence of metabolic syndrome, are beginning to shift this trend[162].

Histologically, esophageal cancer is divided into two subtypes, adenocarcinomas and squamous cell carcinomas[160]. Adenocarcinomas are tumors that arise from mucus producing cells while squamous cell carcinomas are derived from squamous cells as their name implies. Squamous cell carcinomas are presently the most common subtype of esophageal carcinomas, making up 84% of cases globally, but in high-income countries in Northern Europe, North America and Oceania, adenocarcinomas dominate and are on the rise in other countries as well[160]. On the other hand, gastric cancers almost exclusively comprise adenocarcinomas (90-95%), with stromal tumors (GIST), neuroendocrine tumors (NETs) and lymphomas making up the bulk of the

remaining 5-10%[163]. Gastric adenocarcinomas are then usually divided into two histological subtypes based on Lauren's criteria: intestinal and diffuse[164]. Intestinal type adenocarcinomas are typically well differentiated and cohesive structures that frequently become ulcers. Diffuse type usually presents without cell cohesion and instead results in thickening of the stomach wall but doesn't form a discrete mass[165]. Gastric cancer can also be classified anatomically as based on proximity to the esophagus. Those that occur in the proximal areas of the stomach including the cardia, fundus and body are collectively classified as cardia tumors while those occurring in the distal areas including the antrum and pylorus are considered non-cardia[160,163]. Though research and epidemiology suggest that there may be different etiologies driving these subtypes they do share a number of similar features and as of yet no good differentiated treatments exist as based histology alone.

The recent changes in incidence for gastroesophageal cancers are predicated primarily on environmental shifts and geographic differences as clinical treatment has made little to no improvement over the intervening years. Complete surgical resection of all components of the tumor, known as radical gastrectomy, is the only current cure for gastric cancer but is not appropriate or doable in all cases[166]. Esophagectomy, meaning the complete or partial removal of the esophagus, is similarly the current standard of treatment for patients with esophageal carcinomas but these procedures are incredibly invasive with high incidence of morbidity, mortality and overall reduction in patients' quality of life[167]. Increasingly, multimodal adjuvant chemotherapy regimens are being incorporated into treatment but meta-analyses of the clinical trials show mixed results in terms of improved survival[168–170]. One meta-analysis in particular found that significant differences exist in the effect of chemotherapy on cancers between Asian and European patients, suggesting that ethnic and environmental variables may play a role, though

this is confounded by cultural differences in clinical care[171]. Historically, different combinations of 5-fluorouracil and its derivatives, cisplatin or other platinum agents, and radiation have generally been used as adjuvant therapies as they increase survival as compared to surgery alone by about 15%[171–174]. Notably this improvement in survival is mostly found in Asian cohorts but not in Western clinical trials[175]. More recently, addition of docetaxel, a taxane, was found to be effective for perioperative chemotherapy, so now the reference regimen for resectable tumors is FLOT, a combination of 5-fluorouracil, folinic acid, oxaliplatin and docetaxel, but long-term addition of docetaxel can lead to more toxicities[168,175].

Heterogeneity in response to conventional therapies in addition to advancement in personalized treatments of other tumors has led to several attempts to differentiate treatment for gastroesophageal cancer patients on a molecular basis, for instance using genetics. Trastuzumab, a human epidermal growth factor 2 (HER2) antibody, has been used in cases where patients have a large number of HER2 amplifications (HER2+) and clinical trials have shown statistically significant but marginal increases in survival compared to chemotherapy alone[172,176,177]. Based on the results of these clinical trials Trastuzumab has been approved to be added to chemotherapy adjuvant care in HER2+ gastric cancer cases but these only occur in about 15-20% of patients, and is seemingly only effective in patients with a very high large number of HER2 copies[168]. Currently, several clinical trials are also underway to test the use of two anti-PD-1 antibody treatments, nivolumab and pembrolizumab across both gastric and esophageal cancers. For most of these trials they are being tested for efficacy as a second or third line of treatment after failure of earlier treatments but preliminary results are promising. Unfortunately, however, it is estimated that $\leq 25\%$ of patients will respond to immune checkpoint inhibitors, thus limiting their overall use[167,175,178]. A number of other clinical trials for specific targets—including EGFR, MET,

PI3K/mTOR and PARP inhibitors—have yielded disappointingly negative results[175,179–185]. While some of these trials show some promise for further treatment tailoring, the biggest challenge to moving forward with patient specific care is the lack of specific molecular targets due to the highly heterogeneous molecular signatures underpinning these tumors.

3.1.2 Molecular Classifications of Gastroesophageal Cancer

Over the past decade, increasing access to gene expression profiling technologies has helped multiple groups investigate the distinct molecular signatures characterizing gastric cancers in large patient cohorts. Work by researchers at the National University of Singapore, using gene expression patterns across 248 Singaporean and 70 Australian patients, identified three gastric adenocarcinomas subtypes, including proliferative, metabolic and mesenchymal[186]. The results of this study showed promise as different subtypes presented differential response to specific treatments, when tested on matched cancer cell lines. A prospective clinical trial sponsored by this group is currently ongoing to investigate the feasibility of genomic-guided treatment based on these proposed subtypes[187]. In a seminal paper released by The Cancer Genome Atlas research group in 2014, molecular analyses integrating data from RNAseq, protein, mutation, and copy number profiles delineated four different subtypes, including Epstein Barr Virus (EBV) positive, Microsatellite Instability (MSI), Chromosomal Instability (CIN), and Genome Stable (GS)[188]. A later study that included a pan-analysis of all the cancers in the gastrointestinal tract (esophagus, stomach, intestines, and colon) revealed a fifth subtype of Hypermutated samples with Single Nucleotide Variants (HM-SNV)[189]. Subsequent studies by other researchers showed that these groups were associated with different prognoses and responses to adjuvant chemotherapy, when the classification was applied to an independent cohort with additional post-resection follow up

data[190]. Moreover, retrospective meta-analysis of 1,552 patients, accrued by four clinical trials, showed that microsatellite instability status could be used as an effective predictor of good prognosis, both in terms of overall and progression-free survival[175,191]. Notably, patients with microsatellite instability had significantly longer overall survival following surgery, with no adjuvant chemotherapy, while patients with microsatellite stable tumors had the opposite response. Work done by the Asian Cancer Research Group (ACRG) in South Korea over the last several years led to identification of four different subtypes, including mesenchymal-like, microsatellite unstable, TP53 active, and TP53 inactive[192–194]. Their subtypes also showed significant, albeit slight differences in prognosis and recurrence for the molecular subtypes identified by their analysis. Because of the open data availability, most of the groups applied their classification schemas to the other cohorts and were able to recapitulate them, with some level of overlap. This suggests that, despite the high molecular heterogeneity of gastric cancer, there are molecularly distinct subpopulations. However, lack of consensus and harmonization across these different classification schemas suggests that the optimal molecular classification of this disease may still be elusive. Moreover, all of these classification attempts are based almost entirely on gene expression data. Indeed, despite including data from mutational profiles and copy number variants, the causal biological determinants of the identified subtypes are still undefined, thus stymying development of targeted therapeutic interventions.

Other than our analyses in the first MOMA paper, no prior work has investigated the master regulator drivers of gastroesophageal cancer. In that analysis, only gastric adenocarcinoma samples could be considered because the esophageal adenocarcinoma cohort in TCGA was too small to produce statistically sound results. Thus, the pancancer version of MOMA only captured the broadest features of the gastric tumor type, failing to reveal a fine-grain genetic and

transcriptional picture of this disease. For this more in-depth analysis, samples from both cohorts were combined, since recent studies have revealed a high degree of similarity between the two[189].

In the work below, I apply an improved version of the MOMA framework in combination with an iterative clustering methodology to the merged gastroesophageal adenocarcinoma cohort of samples from the TCGA. My results identify 15 molecularly-distinct subtypes that more effectively capture the genetic and transcriptional heterogeneity of this disease, illuminate critical combinatorial biological processes upstream of key MRs, and harmonize across other published datasets. I propose that this novel, molecular-level taxonomy of gastroesophageal cancer, when combined with the elucidation of novel MR-based dependencies and with the subtype-specific drug predictions by the Califano lab's CLIA certified algorithms OncoTarget and OncoTreat[130] algorithms, will help to increase the application of precision medicine approaches in individuals with gastroesophageal cancer.

3.2 Results

3.2.1 MOMA Regulator Ranking and Iterative Clustering

The analyses of this work build off of our previous MOMA framework but include a number of key improvements[85]. All raw data was acquired from the TCGA FireBrowse data repository and processed in the same manner described in the MOMA paper. Briefly, gene expression profiles from 462 STES (stomach and esophageal adenocarcinoma) patients were first transformed to protein activity using the VIPER algorithm using an interactome built on the same samples[63,75]. For the main set of analyses, I generated an internal signature for the VIPER transformation, i.e. I designed the analysis to accentuate differences between the STES samples themselves versus comparing them to all the samples across the TCGA as was done in the first version the analysis. This has been shown in previous VIPER analysis to better highlight subtle subtype differences[64,143]. I then identified global candidate MR proteins as described previously, by Fisher's integration of p-values for (a) their VIPER-measured activity, (b) functional genetic alterations in their upstream pathways, by DIGGIT analysis[83], and (c) additional structure and literature-based evidence supporting direct protein-protein interactions between TRs and proteins harboring genetic alterations, via the PrePPI algorithm[84,85]. The vector of integrated -Log10 p values (Global MOMA Scores) were then used as weights for each TR during the clustering step. This was done with the intention of giving more weight to TRs that had more supporting evidence that they would be drivers as based on calculated associations with candidate genomic events.

For this analysis I also implemented an adapted version of the iterClust framework to interrogate the more granular subtypes present in the STES cohort[195]. Large differences may exist in a cohort of samples that can mask more subtle but significant variations when traditional single pass clustering methods are applied. Repeated clustering within initially identified groups can thus

be informative towards revealing these differences once stronger features of division are already accounted for. The workflow proceeded in the following manner:

Input: STES VIPER matrix weighted by the Global MOMA Scores for each TR as calculated across the cohort

1. Iteration n start
2. Generate clustering solutions for the samples of interest using PAM (partitioning around mediods) clustering for k of 2-9. The distance between each sample was calculated using pairwise Pearson correlation.
3. Select the best clustering solution as based on the highest average silhouette score across all samples.
4. Remove samples determined to be outliers (having a silhouette score < 0.10) in order to promote finding true, high quality clusters.
5. Repeat steps 2-4 within each cluster as determined by the n-1 iteration until all clusters have stabilized and are determined to be homogenous.
6. For every sample that was determined to be an outlier across the analysis, test its similarity to each of the resulting clusters. If its new silhouette score for one of the clusters is above 0.10 then it is added back in to that cluster.

While this was the overall method for the clustering analysis, certain parameters were used to ensure high quality homogeneity and stability of clusters throughout the process. Specifically, for Step 3 when determining the "best" clustering solution for all iterations, a new clustering solution was only selected if the average silhouette scores of the samples was higher than 0.20 to ensure high quality clusters. Additionally, if any clustering solution resulted in clusters that all had fewer than 10 samples it was not selected as I reasoned that having small clusters would artificially inflate

the resulting average silhouette scores as there would not be sufficient samples to draw comparisons from. Lastly, after retesting outlier samples in Step 6, all samples that did not have an individual silhouette score of > 0.10 for any of the clusters were determined to be full outliers and not added back in.

Using this method resulted in 15 stable clusters across 221 samples before outlier re-clustering and 362 after 141 were added back in. Though this resulted in 100 samples from the cohort being removed as outliers, this could be because they are too unique to have a sufficient number of samples similar to them. Analyses of the outliers showed no specific clinical enrichment of any features as compared to the rest of the samples, suggesting it was not a specific feature collectively driving their difference from the rest of the samples. In order to assess the quality of the clustering, I used the sample labels from the recent PanGI analysis from the TCGA, as described previously[189]. Briefly, they are Epstein Barr Virus Positive (EBV), Microsatellite Instable (MSI), Hyper-mutated Single Nucleotide Variant (HM-SNV), Chromosome Instable (CIN), and Genomically Stable (GS). While these are not a pure gold standard, they delineate key biological patterns throughout the cohort and I wanted to determine if my unsupervised clustering methodology, which was agnostic to these labels, recaptured this biology. In addition to these 5 labels, I also labelled samples as based on the degree of their HER2 amplification as this was also a biological phenotype of interest. Currently, Trastuzumab therapy for HER2+ gastroesophageal tumors is the only biomarker driven treatment, but there is a clear heterogeneity in patient response to the therapy, so I wanted to investigate the distribution of these tumors within my clusters as well.

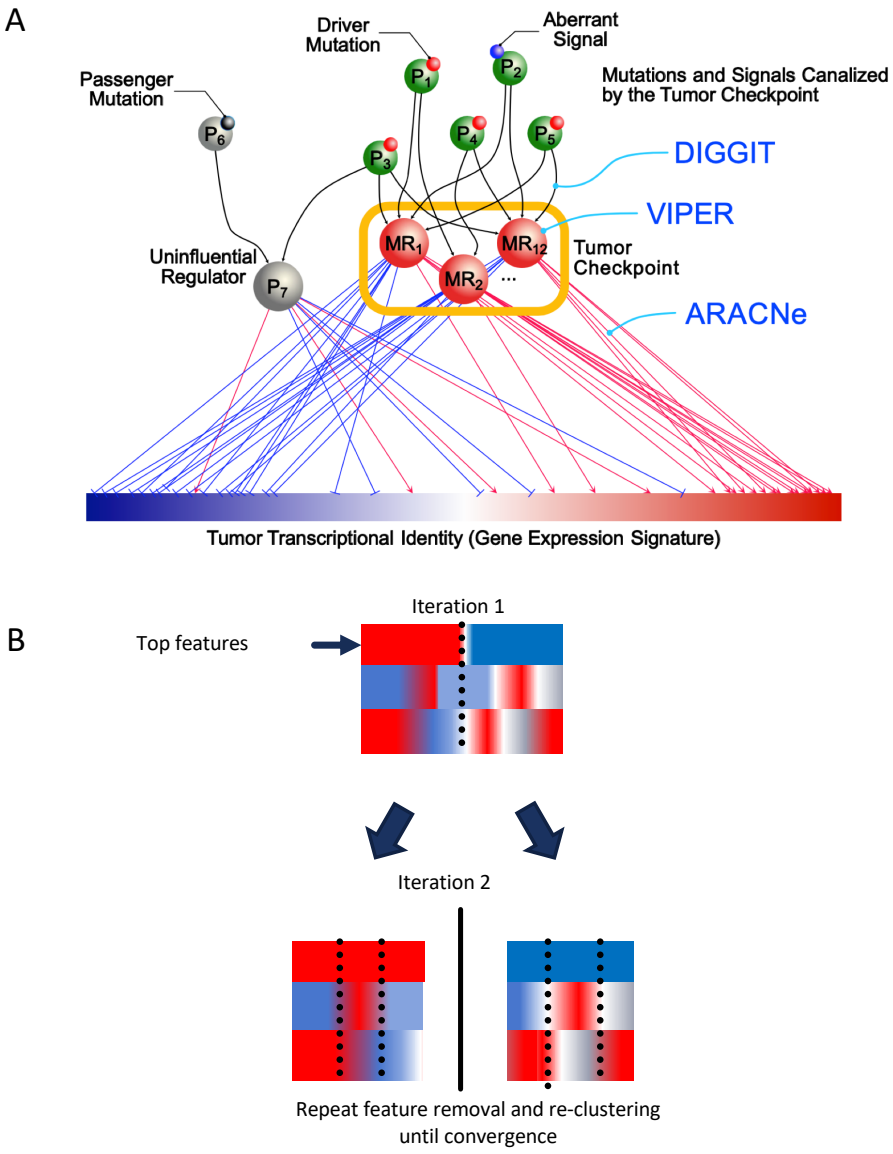


Figure 3.1 MOMA Methodology Overview
(A) General schematic for the integrated MOMA methodology and the bottleneck hypothesis. See Figure 2.1A for full description. **(B)** A schematic for the iterative clustering, showing the influence of certain features at different iterations.

One of the benefits of the step wise progression of this type of clustering is that I could track the driving differences for each iteration, similar to hierarchical clustering. As seen in **Figure 3.2**, each iteration promoted increased separation of the labels of interest. Iteration 1 split the cohort into 2 clusters with C1 having most of the highly amplified HER2 samples as well as the CIN

and MSI samples. C2 on the other hand contained almost all of the GS and EBV labelled samples (**Figure 3.2A**). The second iteration split C1 into two clusters, C1.1 and C1.2, in this case separating into clusters dominated by CIN and a mix of CIN and MSI respectively. C2 was divided into 3 clusters, C2.1, C2.2 and C2.3, effectively putting all of the GS samples in C2.1 and all of the MSI samples in C2.2 (**Figure 3.2B**). Two more rounds of subsequent clustering led to a stabilization of 15 clusters with significant stratification of a number of different TCGA labels across them. (For simplicity final subtype clusters will henceforth be referred to using S and their final cluster number, from 1-15 and not using the nomenclature including their parent cluster, ie C1.1.1.1 will instead be S1). See **Figure 3.2** for final subtype delineations and **Table 3.1** for specific enrichment statistics.

Table 3.1 Final Subtype Summary Statistics

Subtype	iterClust name	Number of Samples	Final Avg Sil Score	Enrichments				
				CIN	HM-SNV	MSI	GS	EBV
1	1.1.1.1	67	0.36	2.8E-08	1	1	1	1
2	1.1.1.2	28	0.12	0.0087	1	1	1	1
3	1.1.2.0	15	0.06	0.018	1	1	1	1
4	1.2.1.1	32	0.32	1	1	1	1	1
5	1.2.1.2	11	0.13	1	1	1	1	1
6	1.2.2.1	18	0.14	1	1	2.9E-04	1	1
7	1.2.2.2	11	0.32	1	1	6.0E-04	1	1
8	2.1.1.2	41	0.20	1	1	1	9.8E-08	1
9	2.1.1.3	33	0.23	1	1	1	1	9.9E-04
10	2.1.2.0	12	0.03	1	1	1	1	1
11	2.1.3.1	44	0.23	1	1	1	0.0031	1
12	2.2.2.0	10	0.25	1	1	1	1	0.0021
13	2.2.3.0	14	0.37	1	1	1.0E-08	1	1
14	2.3.1.0	16	0.14	1	1	0.018	1	0.16
15	2.3.2.0	10	0.23	0.14	1	1	1	1

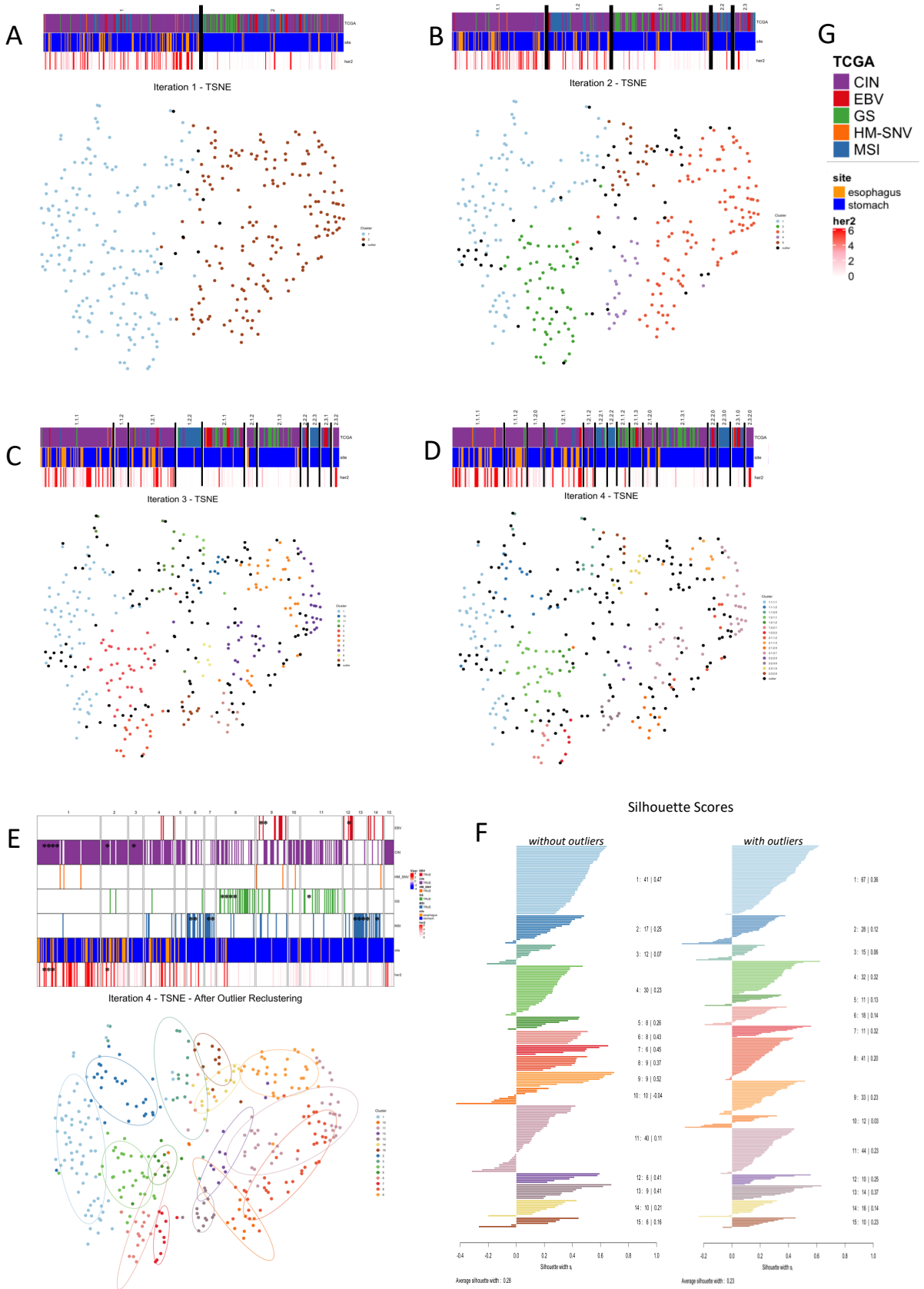

A
Iteration 1 - TSNE
B
Iteration 2 - TSNE
G
TCGA
CIN
EBV
GS
HM-SNV
MSI
site
esophagus
stomach
her2
6
4
2
0
C
Iteration 3 - TSNE
D
Iteration 4 - TSNE
E
Iteration 4 - TSNE - After Outlier Reclustering
F
Silhouette Scores
without outliers
with outliers

Figure 3.2 Results of Iterative Clustering
(figure on previous page). **(A-E)** TSNE plots of samples colored by their cluster assignment for that iteration. Bars across the top show TCGA classification labels, tumor tissue location and HER2 amplification respectively. **(F)** Silhouette scores for the final clustering solution both without outliers and with them added back in. **(G)** Color legend for panels A-E. **** $p < 0.00001$; *** $p < 0.0001$; ** $p < 0.001$; * $p < 0.05$

In addition to being enriched in a number of previously identified biological labels, the resulting subtypes also had a number of notable differences in survival. Comparison of survival across all 15 clusters at once did not reveal significant differences ($p = 0.21$ by Kaplan-Meier) (**Figure 3.3A)** in part because of the number of subtypes and because this cohort has been notoriously hard to stratify based on survival due to limited outcome data. Looking at specific subtype pairs does illuminate some significant differences. Specifically, comparing S6, the subtype with the best survival, to the two subtypes with the worst survival, S3 and S12, results in significant survival separation ($p = 0.001$ and $p = 0.0098$). Additionally, comparing S6 to the two subtypes enriched in GS samples, S8 and S11 also revealed significant separation ($p = 0.047$) which is concordant with the fact that patients who have GS tumors tend to have the worst outcomes.

3.2.2 Tumor Checkpoint MRs

As defined in the original MOMA paper, a Tumor Checkpoint is a module with the minimum MRs necessary to implement a tumor's transcriptional identity by canalizing genomic events in its upstream pathways. In order to calculate this, we used saturation analysis as previously described to refine the ranked list of regulators (based on their MOMA scores) to the subset necessary to optimally account for each subtype's genetic landscape. In the first version of MOMA the same Global MOMA ranking was used across every subtype, thus minimizing the effect a small number of genomic events might have, even if they are highly enriched in a particular subtype. To address

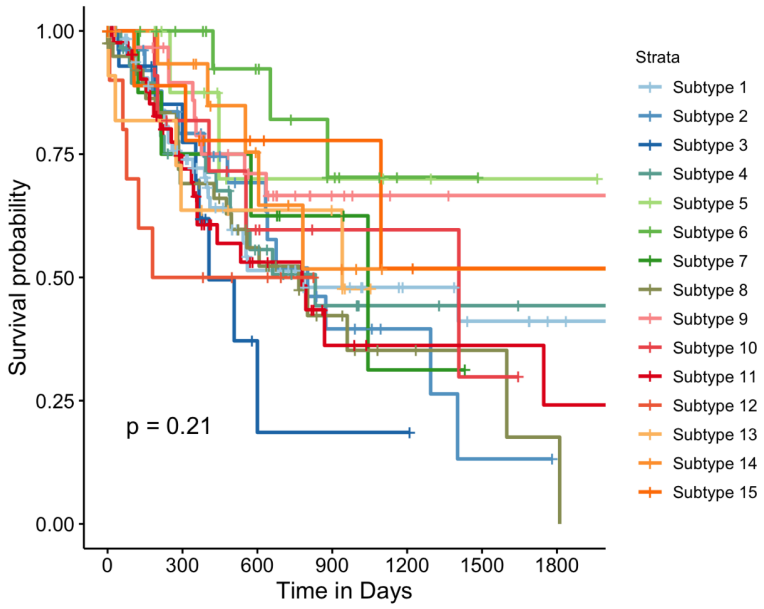


Figure 3.3 Survival Curves for the Final 15 Subtypes

this, I improved upon the original framework by re-ranking the candidate TRs on a subtype by subtype basis to generate a Subtype Specific MOMA ranking to use for the saturation analysis. This was done by taking the resulting subtypes from the clustering analysis and repeating the integrated ranking methodology, but only considering genomic events that were over represented in that subtype as compared to the cohort as a whole. Subtype specific events were considered to be over represented only if they occurred in more than 2 samples in that subtype and if they had a $p > 0.50$ after doing a proportions test and using FDR for multiple hypothesis correction. Using this methodology to re-rank TRs prior to the saturation calculation in combination with the iterative clustering improved the quality of the tumor checkpoint analysis by increasing the overall saturation percentages captured across the cohort (**Figure 3.4**). In the first version of the analysis, the saturation percentages for the 3 clusters ranged from 4.0% - 21.4% (average = 13.7%). In the new analysis the range was 13.2 - 80.0% (average = 44.7%) across the 15 clusters, a marked

improvement. In this way, the analysis more accurately captured the different upstream biology of each more granular subtype versus focusing only the largest signals across the whole cohort.

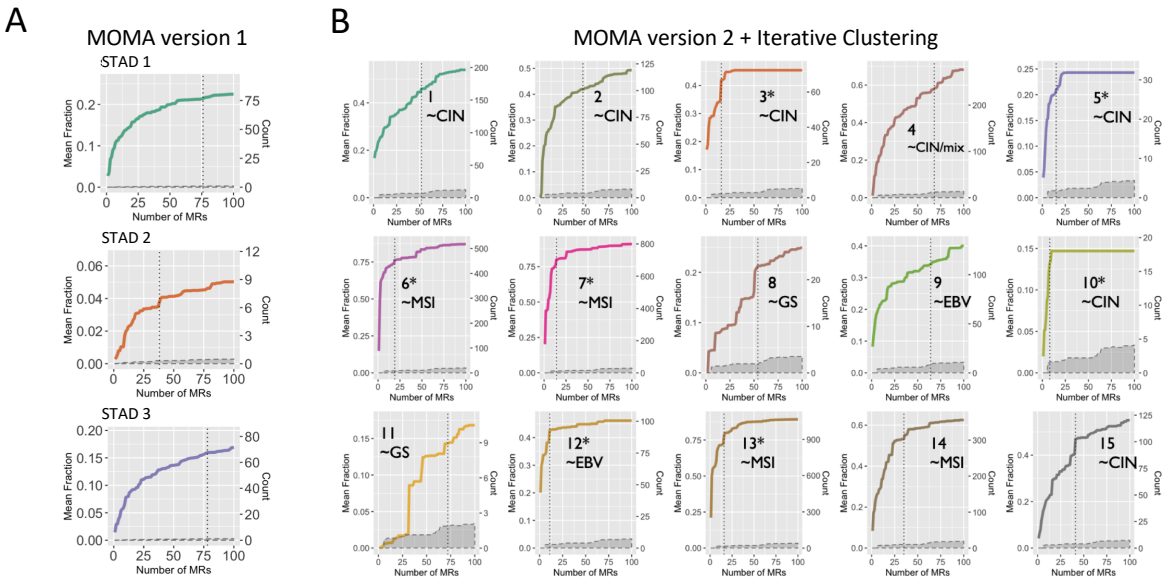


Figure 3.4 Genomic saturation analysis of candidate master regulators.
As described in Figure 2.3, Individual curves show the average fraction of functional genomic events in each sample identified upstream of the top *n* MOMA-inferred MR proteins for each subtype, as *n* increases from 1 to 100. Saturation curves produced by the null-hypothesis—i.e., *n* randomly selected MRs from 1,253 non-statistically significant regulatory proteins (i.e., the bottom half of all MOMA-ranked proteins)—are shown in gray. Vertical dashed line indicates the saturation threshold. **(A)** The results from the first MOMA analysis vs **(B)** the results of the present analysis.

To further validate the biological quality of these predicted checkpoint MRs (cMRs) I then tested to see if they were enriched in essential genes as based on the Achilles Project data[113]. I used a workflow similar to the one detailed in the original MOMA analysis but with updated Achilles results from their Quarter 1 2021 release, see **Figure S2.5** (in Appendix A). Briefly, I took the VIPER protein activity profiles for each sample and used Stouffer's Integration for each TR across each subtype to generate a single integrated VIPER profile to represent each subtype. Gene expression profiles from all available cell lines from the CCLE were then transformed into VIPER

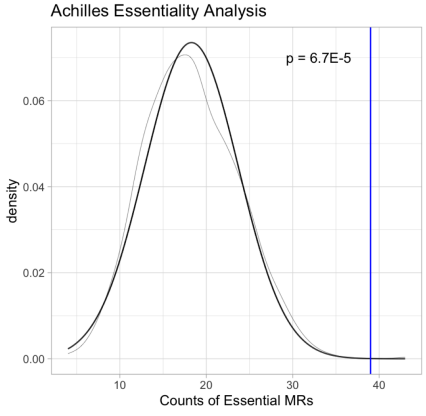
protein activity profiles and viperSimilarity was used to calculate similarity scores between each of the 15 integrated subtype profiles and each cell line. The top and bottom 25 TRs were used for the similarity scoring as this has been shown to optimally find matches based on relevant biological signal and not artifacts from cell lines or other models. Cell lines were considered a match if their similarity score was $p < 0.01$ and non-matches for the null model were selected as those with a $p = 1$, after Bonferroni correction. Relative essentiality in the matching cell lines as compared to the null cell lines was computed by comparing the essentiality rank of the cMR of interest in the cell line matches as compared to the rank of the MR in the non-matching cell lines, using a non-parametric rank-based Mann-Whitney-Wilcox test. In total 39 cMRs were determined to be significantly relatively essential subtype-specific MRs after multiple hypothesis correction (Benjamini-Hochberg FDR < 0.05), and this total amount of essential MRs was significant with a $p = 6.7 \times 10^{-5}$ after fitting a null distribution to 10^4 random selections of the same number of TRs for each checkpoint. Essential MRs are highlighted in **Figure 3.5B** and bolded in **Table 3.2**.

Figure 3.6A shows the full plot of all the VIPER activities for the cMRs for each of the 15 resulting subtypes. From this global perspective a number of patterns emerge. Almost all of the cMRs from subtypes 1-7 are highly active in these subtypes but have middling to low activity in subtypes 8-15. The converse is true of the cMRs for subtypes 8-11. This is consistent with the division of samples in the first iteration of clustering. Amongst these larger differences in VIPER activities, more granular ones can be seen at this level and will be discussed more in the following sections. As an initial accounting for the biology and quality of these subtypes I used the labels from the TCGA analysis to look for enrichments. In particular, as can be seen in **Table 3.1**, MSI samples were primarily enriched in subtypes S6, S7 and S13. Additionally, GS samples were

Table 3.2 Checkpoint MRs for each subtype

subtype	Checkpoint MRs
1	CDKN2A DNMT3B RNF2 TAF4 NONO SETDB1 COPS5 YY1 **CHAF1B** HDAC2 MED20 TAF2 TONSL ETV4 POGK ACTR5 TARDBP TCFL5 TGS1 ZFP64 **GTF2IRD1** CDC5L MORC2 TRIM28 SUMO1 KAT2A BAZ1B HLTF PLAGL2 NELFCD ING5 **TP53** TRIM27 MCM8 ZC3H8 SUDS3 SRSF10 ZBTB9 KDM1A PRMT5 RBBP7 BMP7 MAPK15 ZNF556 SUPT5H OTX1 YEATS4 PROX1 RUVBL1 TFDP1 TOP2A PPP1R10
2	DNMT3B HLTF NR5A1 ARID3A PLAGL2 NR6A1 ZNF556 ZNF765 CHD7 DNMT3A TAF2 AHCTF1 ZNF544 SCML2 ASXL1 ZNF749 **TBX4** ADNP CAND1 SUPT16H **MCM8** ING5 HOXC13 ZNF217 ZBED4 ZNF572 BMP7 ZNF551 ZNF697 SMARCC1 GZF1 ZNF473 ZNF480 ERCC3 HOXD12 DLX4 BMI1 NFXL1 CDC5L AGO2 HCFC1 CTNNB1 ZSCAN22 CHD4 ZNF133 POU4F3 TARBP1
3	NR5A1 HOXC13 HOXC12 ANKRD1 ZNF280A SOHLH1 ZNF114 ALX1 MTA3 SSX4 ESX1 SSX5 TGIF2LX PCGF2 SSX1 MNAT1
4	SETDB1 PLAGL2 MYC TAF4 RNF6 **COPS5** RNF2 TAF2 NONO NR6A1 RNF4 **CDC5L** MCM8 HOXA11 CDK8 TRIM27 CHD7 **TAF5L** EMX1 CAND1 OVOL1 EAF1 YY1 SMARCC1 SATB2 TGS1 LRPPRC PRMT5 CHAMP1 TARDBP HOXA10 ING5 CDH1 DDX1 NFX1 ZBED4 SFMBT1 TFB2M CNOT7 TARBP1 SUDS3 ADNP CTBP2 **KDM1A** ZDHHC13 ZDHHC23 NFXL1 RBBP4 ZNF3 **RB1** SUPT16H **TADA1** CTCF ZIC5 NOLC1 XRN2 GTF2H3 ZC3H8 NELFCD ACAD8 ACTR5 KDM4C ZNF664 SLC30A9 FOXD4 **KLF5** GMCL1 ZNF343
5	ERN2 CASZ1 PPARA ATP8B1 ARX RFX6 **FOXA1** TRAK1 RORC **CASK** LPIN2 ZNF710 FEV ZNF774 NPAS2
6	COPS5 EMX1 ELP3 GTF2E2 C1QBP SUMO1 **ASH2L** SNW1 MLX MYC TRIM27 **TP53** HDAC2 **HOXA10** HOXA11 XRCC6 ERN2 SIRT7 PPP1R8
7	MYC ELP3 CDC5L GTF2E2 SNW1 C1QBP CNOT7 EMX1 RBBP4 XRCC6 PRMT5 TGS1 ATF1 ZFAND1
8	TRAK2 CREBL2 GATA1 ZNF536 STAT3 SAP30L ZMAT3 AFF3 NFIB SMAD2 ZNF671 ZSWIM6 SMARCA2 KLF11 ADH1A IGF1 ZEB2 STAT5B ZNF727 FOXO1 ELK3 FOXN3 CD86 ZIK1 SPI1 KAT2B ZNF483 SETD7 CREBRF ESRRB ABCG1 ELP2 EBF2 HCFC2 VAV1 ESR1 NFKB1 ZNF660 SOX8 IKZF1 ZBTB38 TFEC ZNF486 GDF7 L3MBTL4 FOXP3 ZNF438 ZNF132 DBX2 SMAD1 NSD3 ZNF24 ZNF561 MNDA
9	STAT1 MOV10 RELB ICAM1 IRF1 MAX ZNF683 CD86 BATF IFNG FOXP3 TNFRSF4 SNAI3 **IRF2** TBX21 **TRAFD1** CXXC1 **NR1H3** MEF2B ETV7 **SPI1** BORCS8-MEF2B STAT5A BATF3 BATF2 FOXB1 **PML** ZBED2 VAV1 IRF5 RFX5 BTG1 GATA3 CIITA **IL10** SLC11A1 NFKB2 HCLS1 KEAP1 **RUNX3** EOMES IKZF3 AKNA LITAF TFEC **NFKB1** MNDA **GFI1** TOX2 WNT1 **IKZF1 MED9** SBNO2 LYL1 SP140 **SAP30 PRDM1** SP100 CEBPE **IRF4** TRIM22 PARP14 ZBTB32 OLIG2
10	SOX21 KLF4 EYA2 RORC PTH NKX6-2 LRRFIP1 RNF141
11	PURA TERF2IP HEXIM1 ZNF394 TCEAL1 **ZFP36L1 SCX** PHF1 TCEAL8 ZNF34 CAVIN1 TCEAL4 HOXA3 GTF2A1L HOXA2 PHF20 HOXA5 HOXB2 HABP4 **TCEAL3** DNAJB6 **TCEAL6** HOXB4 PBXIP1 HOXA4 EID2B SIX2 MEIS2 TEAD3 ZNF358 PNRC1 SMAD1 ISL2 EID1 SIRT4 TCEAL2 APP RBPMS SCMH1 NPAS4 ZCWPW2 FOXP1 TSC22D1 RFX2 MEAF6 PRRX1 CAMK2A **EMX2** ZNF615 ZFHX3 THRB MKX HAND2 DLX1 LMO1 ZSCAN31 BARX1 KLF10 ID4 PIAS3 PKNOX2 NFYB MXD4 **SMARCD3 TCEAL7** SOX2 PRDM8 MEIS1 MMP14 TSHZ1 MAEL GLI2
12	STAT1 IRF1 MOV10 IFNG RFX5 **TRAFD1** CD86 RBBP4 ICAM1 ZNF683 ZNF317
13	GTF2E2 ELP3 SNW1 PRDX3 C1QBP XRCC6 SMAD2 ZFAND1 NRBF2 RCHY1 RBBP4 RB1 CNOT7 ING3 EAF1 ZIC5
14	C1QBP TAF9 TRIM28 PCBD1 ELOB GTF2A2 PPP1R8 TSFM NEDD8 SAP18 APEX1 EDF1 PHB2 MED28 UBE2I TFB1M PA2G4 ELOF1 PRDX3 RFXANK ENO1 ZNHIT1 PDLIM1 RUVBL2 PHF5A SRSF2 ZNF511 MRPL12 HNRNPAB SLIRP SUV39H1 HMGB1 ABT1 MCM5 MLX
15	GLMP CERS2 PRPF6 ELK1 ATF6B CTBP1 FOXP4 DEDD MED8 SIRT2 MED29 ZBTB45 PREB VPS25 CREB3 ZSWIM3 TAF6 ZSWIM1 SMAD6 SNAPC5 MESP1 RXRB FIZ1 ZNF768 THRA MAF1 ERF ZNF784 KAT8 SCYL1 LRCH4 IRF2BP1 CRTC2 HDAC11 ZNF444 MESP2 ZFPM1 RELA USP21 NFYC CREB3L4

A



B

Subtype	Essential cMRs
1	CHAF1B GTF2IRD1 TP53
2	TBX4 MCM8
3	
4	COPS5 CDC5L TAF5L KDM1A RB1 TADA1 KLF5
5	FOXA1 CASK
6	ASH2L TP53 HOXA10
7	
8	
9	IRF2 TRAFD1 NR1H3 SPI1 PML IL10 RUNX3 NFKB1 GFI1 IKZF1 MED9 SAP30 PRDM1 IRF4
10	
11	ZFP36L1 SCX TCEAL3 TCEAL6 EMX2 SMARCD3 TCEAL7
12	TRAFD1
13	
14	
15	

Figure 3.5 Relative Essentiality of cMRs.
(A) Total count of subtype cMRs found to be relatively essential (blue line) compared with the probability density generated by 10^4 random selections of the same number of cMRs and fitted to a normal distribution to assess statistical significance. **(B)** Subtype specific relatively essential MRs.

highly enriched in both subtypes S8 and S11. Moreover, almost all of the highest HER2+ samples were disbursed between subtypes 1-4, with S1 and S2 having the most significant enrichment. All of these suggest that my new MOMA based classification is both identifying previously identified biological features as well as revealing novel ones.

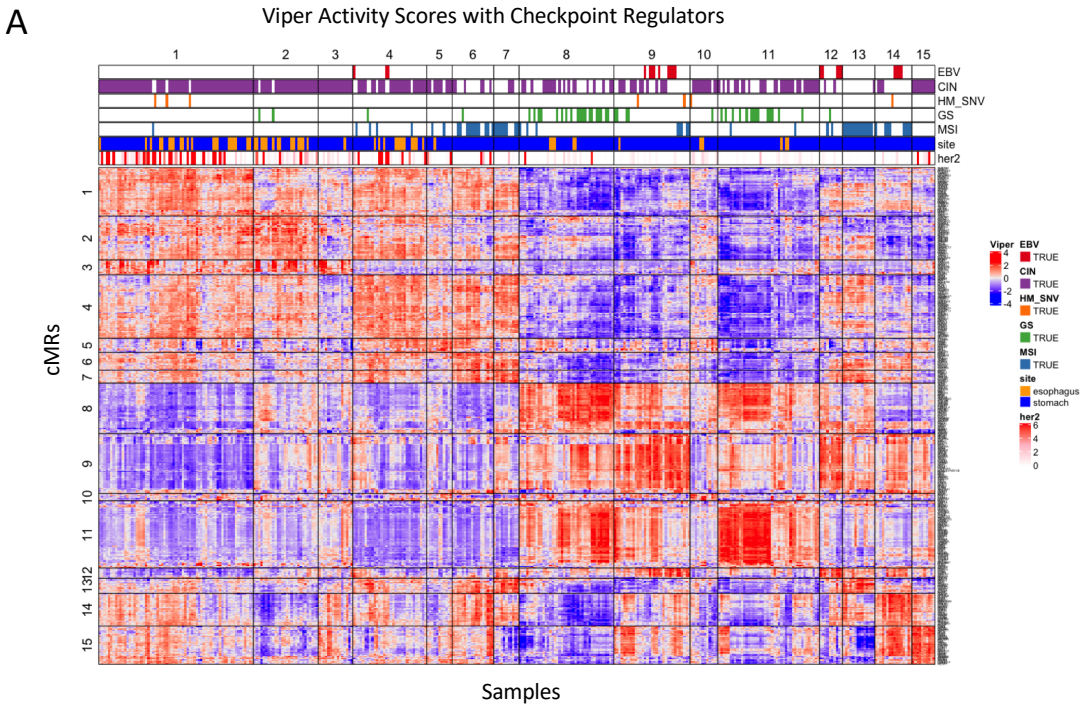


Figure 3.6 VIPER Activity Scores across the STES cohort.
(A) MR-based clustering heatmap for the STES samples. Rows represent checkpoint MRs while columns represent individual samples. Color scale is proportional to protein activity (red activated; blue inactivated). Top annotations are the same as Figure 3.2G, and the same throughout.

3.2.3 Microsatellite Instable Subtypes: S6, S7 and S13

The first set of subtypes of interest in this analysis were those enriched in MSI samples: S6, S7 and S13 ($p = 2.9x10^{-4}$, $p = 6.0x10^{-4}$and $p = 1.0\ x10^{-8}$ respectively). MSI is a well characterized phenotype and these samples clustering consistently together served as a good positive control that

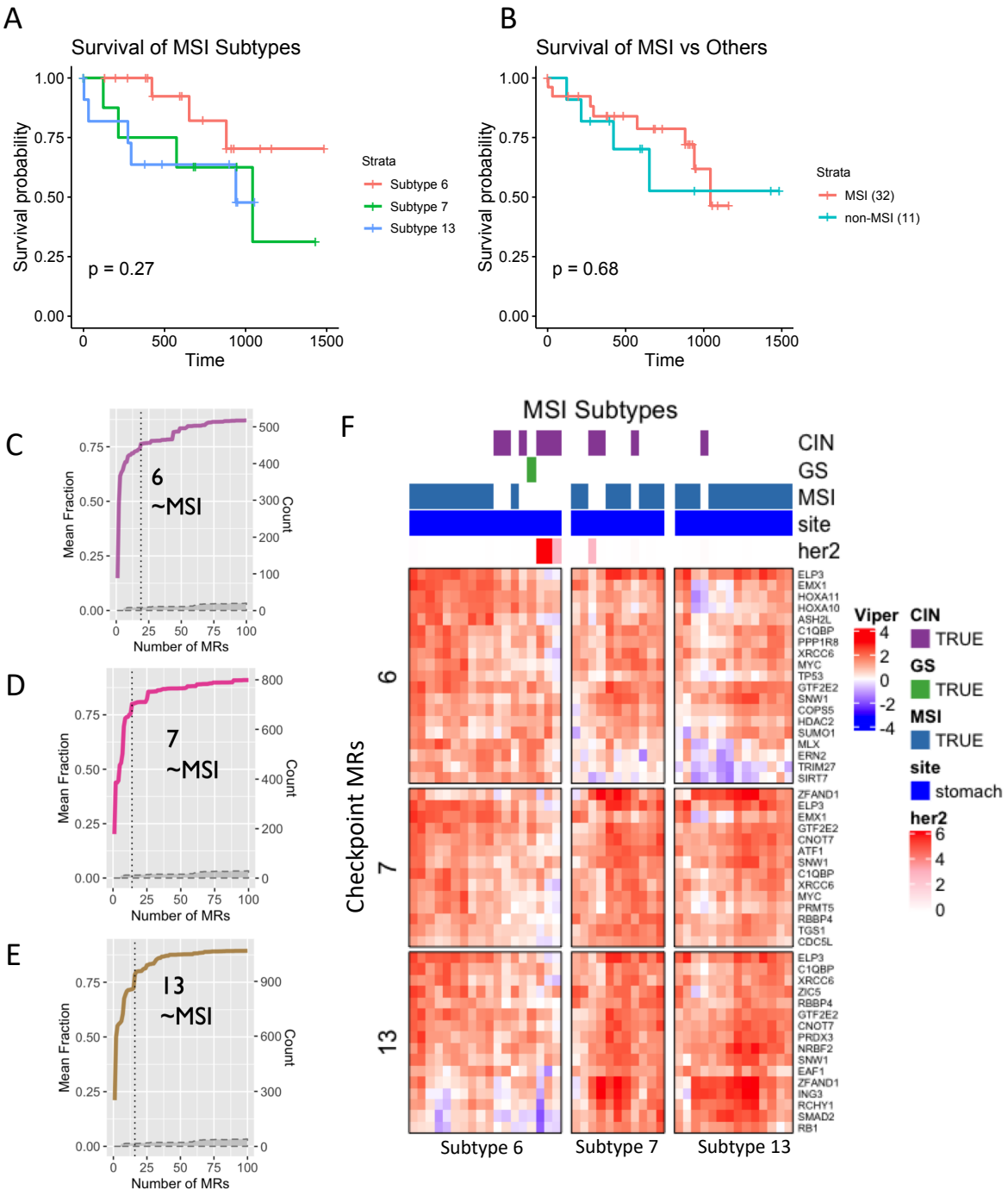


Figure 3.7 Summary of MSI Subtypes.
(A) Survival curves for all 3 MSI enriched subtypes. **(B)** Survival of MSI samples across all three subtypes vs the non-MSI samples. **(C-E)** Saturation curves as described in Figure 3.4. **(F)** Heatmap of VIPER activities of the checkpoint MRs as described in Figure 3.6.

the methodology was effectively capturing known subtypes[189,194]. While across these three there were no significant differences in survival, S6 was the best surviving subtype out of all 15, which is consistent with the literature and clinical findings showing that MSI high patients tend to fair better overall[175,191]. Additionally, analysis of non-MSI samples across these clusters did not reveal significant survival differences from the MSI samples.

Saturation of upstream genomic events occurred at 76.3%, 80% and 78.4% total events respectively for each subtype (**Figure 3.7C-E)**. These were some of the highest rates of saturation across the analysis, likely because of the overall high frequency of genomic changes. All three of the subtypes had high incidence of amplifications across chromosome 8, though for S13 predicted driver events were identified only on the q arm. Interestingly all of the non-MSI samples in S6, including a GS sample, had many of the same amplifications. A number of genomic events appeared as drivers at a high frequency across all three subtypes, most of them related to chromatin remodeling (**Figure 3.8)**. These included ARID1A (39%-86%), NBN (64-72%), and members of the KMT2 and CHD gene families. Though KMT2D and KMT2C were present across all three subtypes, KMT2A and KMT2B were not found to be drivers in S7. CHD4 was a driver in S6 and S13 but not in S7 while mutations in CHD7 appeared in S7 and S13. Overall S13 had the highest overall frequency of different point mutations across the three subtypes. Absent across the drivers of all the subtypes were either of the canonical MSI related genes, MLH1 and MSH2. This is likely due to the fact that in cancers of the upper gastrointestinal tract these genes are more often affected via epigenetic silencing and not mutation[189].

Gene set enrichment of Reactome Pathways present across the upstream drivers showed a high degree of similarity between S6 and S13. In particular both had significant enrichment in genomic events related to Diseases of glycosylation, ECM proteoglycans, and Interaction between

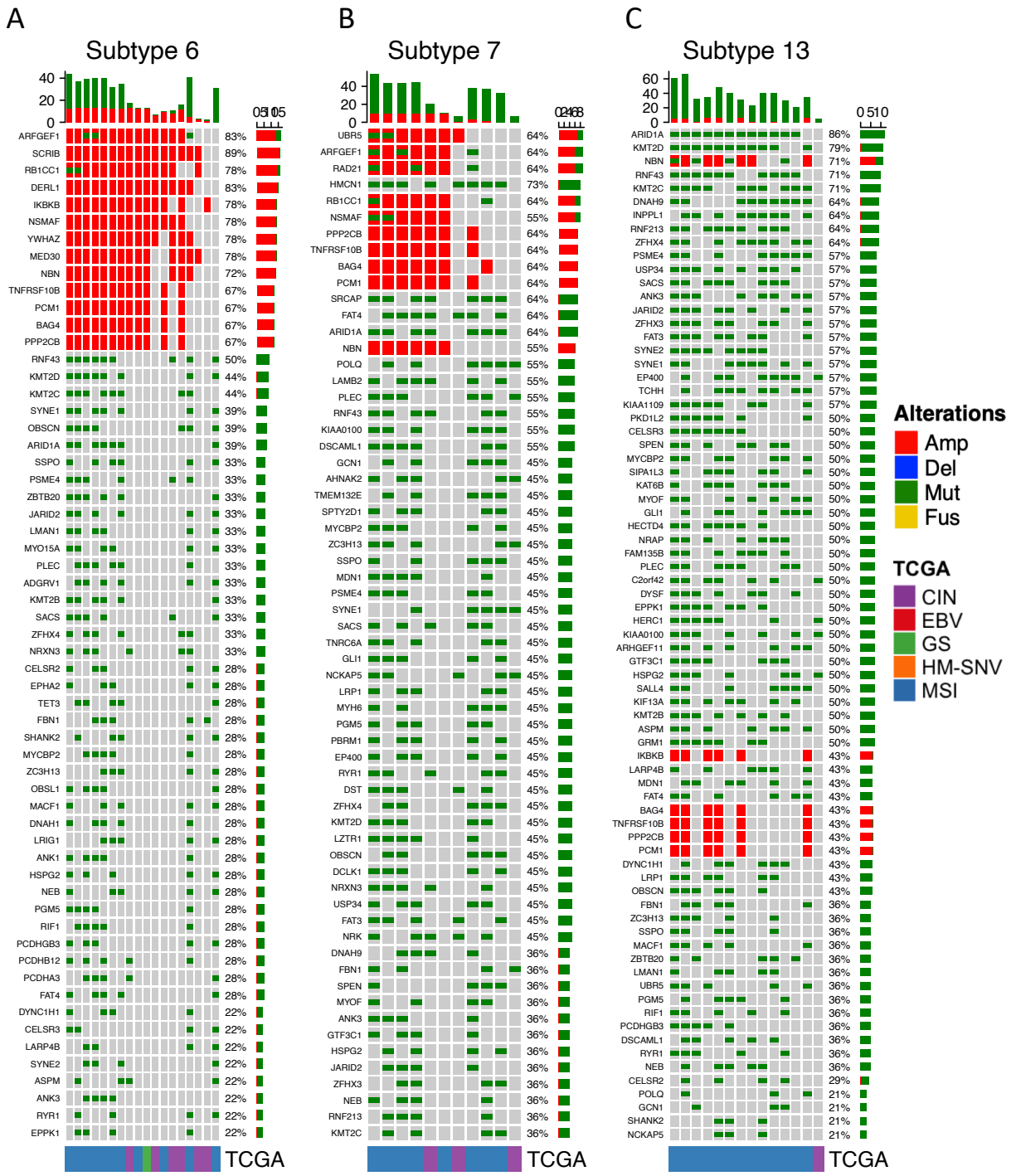


Figure 3.8 OncoPrint plots for MSI subtypes.
OncoPrint plots showing predicted driver events per sample upstream of the cMRs for **(A)** S6, **(B)** S7 and **(C)** S13. Horizontal histograms and percent numbers show the fraction of samples harboring the specific event type. Vertical histograms show the number of events detected in each sample.

Figure 3.9 Reactome gene enrichments of MSI subtypes.
Gene set enrichment of upstream drivers, top 100 MRs and top cMR targets respectively. Color of each dot corresponds to its p value (red being more significant) and the size corresponds to the gene ratio for each gene set.

L1 and Ankyrins, all pathways related to cell membrane structure and connectivity to the ECM, processes particularly important to cancer progression and morphogenesis (**Figure 3.9)**. While genomic events upstream of S7 were not enriched in these processes, it had significant enrichments for Laminin interactions, MET promotes cell motility and Assembly of collagen fibrils and other multimeric structures, processes also related cell structure and interactions with the ECM but via different mechanisms. This convergence of phenotype via different pathways suggests that while these subtypes are similar they may have progressed to this point by different genomic alteration cascades. Notably both S6 and S7 were enriched in samples that were significantly more stem cell-like ($p = 0.0024$ and $p = 0.00047$ respectively by Students T-test) (**Figure 3.10**)[135,189]. Additionally, S6, the more stem cell-like of the two, had significantly fewer leukocytes, an association that's been shown across a number of cancers ($p = 2.8x10^{-5}$ by Student's T-test)[196].

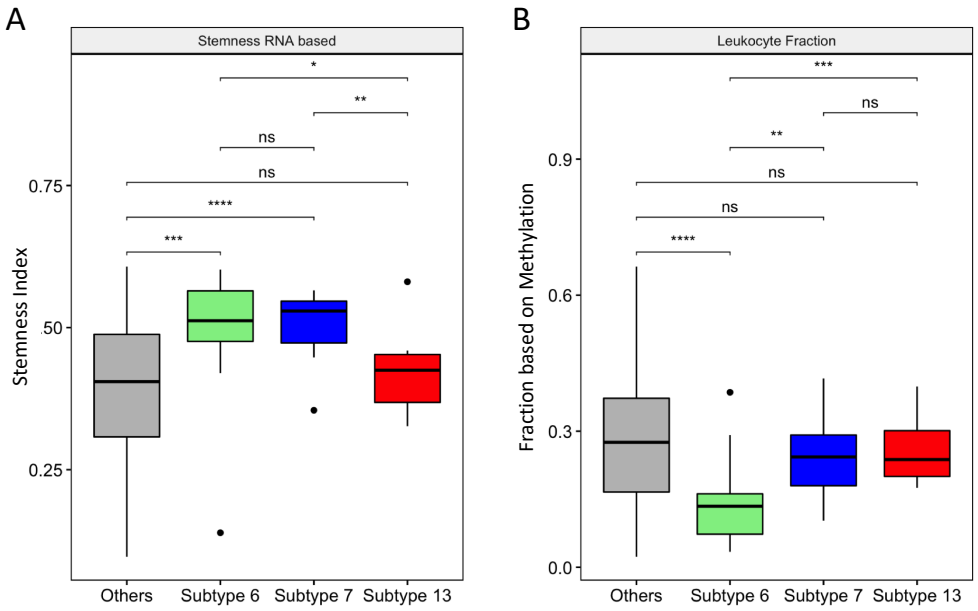


Figure 3.10 Phenotypic features of MSI subtypes.
(A) Box plots of relative stemness between each subtype vs all others**. (B)** Box plots of leukocyte fraction. Both as reported in [189]. ns: p > 0.05, *: p <= 0.05 , **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001.

Analysis of the cMRs for these subtypes also showed a fair amount of similarity. Four cMRs occurred across all three subtypes, GTF2E2, ELP3, SNW1, XRCC6, and C1QBP **(Figure 3.7F)**. The first two are components of the RNA polymerase complex, mediating initiation and elongation respectively, while SNW1 can serve as a transcription coactivator at certain Pol II promoters. XRCC6 is a helicase that is involved in DNA non-homologous end joining, particularly after double stranded breaks. C1QBP is a multifunctional protein that is involved in a number of processes including inflammation, ribosome biogenesis, regulation of apoptosis and mRNA splicing. Though an analysis of activity of all three sets of cMRs across all the subtypes showed broadly high activity across all three, certain cMRs had high specificity to their associated subtype. In particular, S6 cMRs TRIM27 and SIRT7 were specifically high only those samples, while SMAD2 and RB1, cMRs in S13 were highest in those samples with moderately elevated activity in S7. A broader analysis of the top 200 most significantly dysregulated TRs across each subtype (after Stouffer integration of sample specific regulator scores) showed similarity between S7 and S13, but not S6 (**Figure 3.9**). Top dysregulated TRs from S7 and S13 had significant enrichment in Chromatin modifying enzymes, SUMOylation, and gene expression pathways mediated by PPARalpha and PTEN. For S6 the only pathways that appeared as highly enriched across the dysregulated MRs were Notch-HLH transcription pathway and Nuclear receptor transcription pathway, both of which were also significantly enriched in S7 and S13.

To further characterize these subtypes, I analyzed the top inferred downstream targets of each of the cMRs ($p < 0.05$) (**Figure 3.9**). Gene set enrichment across these again showed largely similar patterns of biology mostly related to cell cycle checkpoints, DNA replication and synthesis, and RNA processing. Interestingly, a number of the pathways enriched across all three were virus related, including HIV infection, Rev-mediated nuclear export of HIV RNA, Interactions of Rev

with host cellular proteins and NEP/NS2 interactions with Cellular Export Machinery. Further inspection showed that the cMRs shared across these subtypes (SNW1, C1QBP, GTF2E2 and XRCC6) were the main regulators of the target genes in these enrichments. This phenotype could be a result of the fact that MSI tumors are constantly making mutant proteins that can become neo-antigens and may appear viral, thus triggering the same pathways.

3.2.4 Genomically Stable Subtypes: S8 and S11

Another set of subtypes of interest in this analysis were the genomically stable enriched subtypes, S8 and S11 ($p = 9.8x10^{-8}$ and $p = 0.0031$ respectively). As with the MSI subtypes, finding that most of the GS samples clustered together was a good positive control for the efficacy of the methodology. These subtypes have been historically hard to categorize because of their limited number of mutational events and are also often the most lethal. This was consistent in my analysis as well, with S8 and S11 being two of the worst subtypes in terms of overall survival. Notably, though the subtypes are mostly GS samples, an analysis of the GS samples vs the others revealed no significant differences in survival, indicating a similarity in outcomes for these patients as well.

Saturation of upstream genomic events occurred at 21.3% and 14.3% total events respectively for each subtype (**Figure 3.11D-E)**. While these values were on the lower end of the saturation range this is to be expected as many of these samples have few mutations overall and some may have been to infrequent to be statistically significant (**Figure 3.12**). S8 was dominated by CDH1, HECTD4, and FCRL3 mutations as well as CLDN18-ARHGAP26 fusions, all in a largely mutually exclusive manner. Additionally, most of the patients with either HECTD4 or FCRL3 mutations also had amplifications in ARFGAP1, UBE2C, POFUT1 and PIGU, all of which are on the q arm of chromosome 20. This suggests that having one of these mutations,

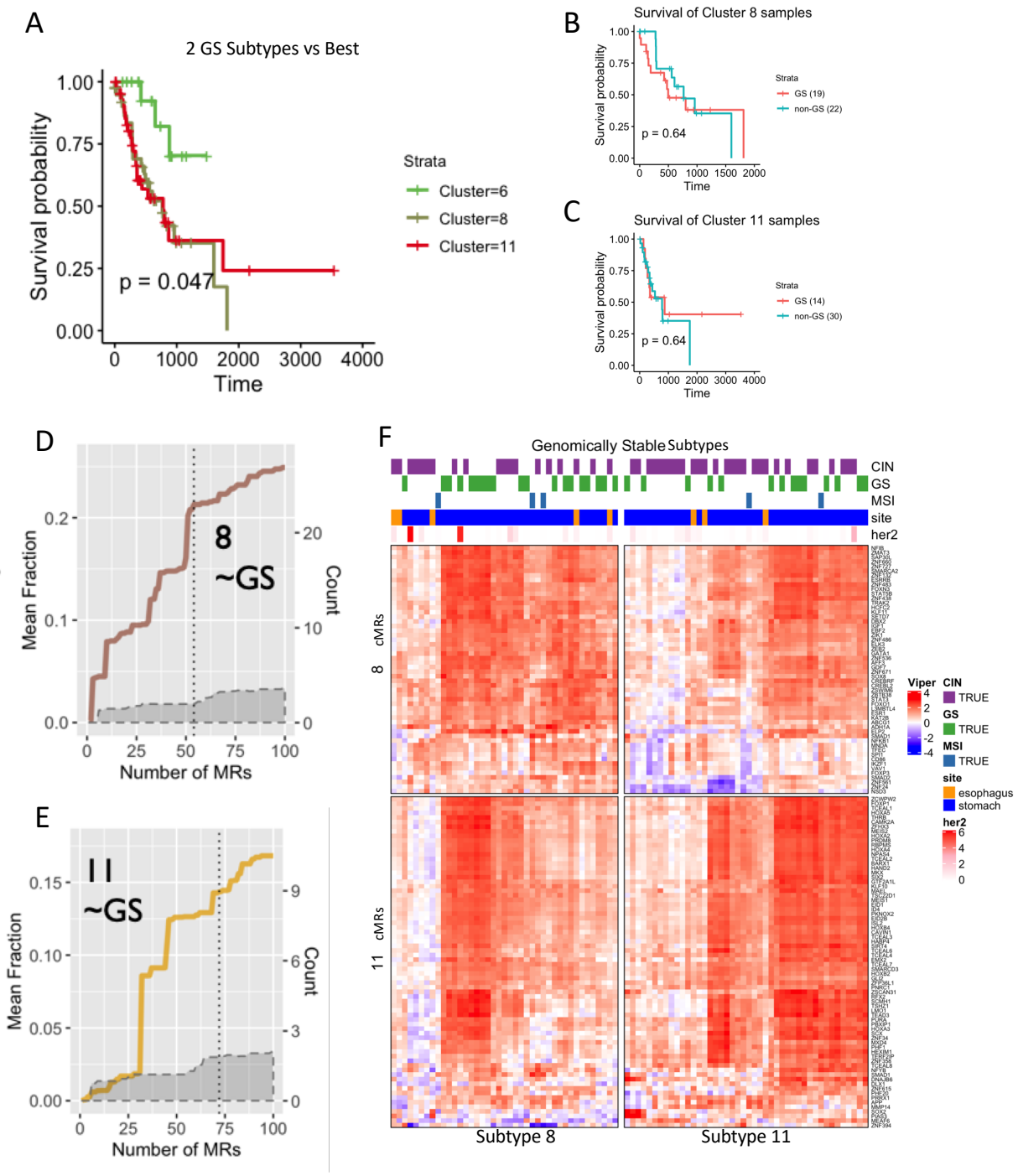


Figure 3.11 Summary of GS Subtypes.
(A) Survival of 2 GS subtypes vs the best surviving subtype. **(B-C**) Survival of GS samples across all both subtypes vs the non-GS samples. **(D-E)** Saturation curves as described in Figure 3.4. **(F)** Heatmap of VIPER activities of the checkpoint MRs as described in Figure 3.6.

potentially in combination with these amplifications is likely enough to implement the downstream cMRs. S11 has a higher number of CIN samples along with the GS samples and its driver events included a number of the same amplified genes along chromosome 20. These samples did not have the corresponding mutations seen in S8 but a few have mutations in CDH1, CPQ and OR10K1.

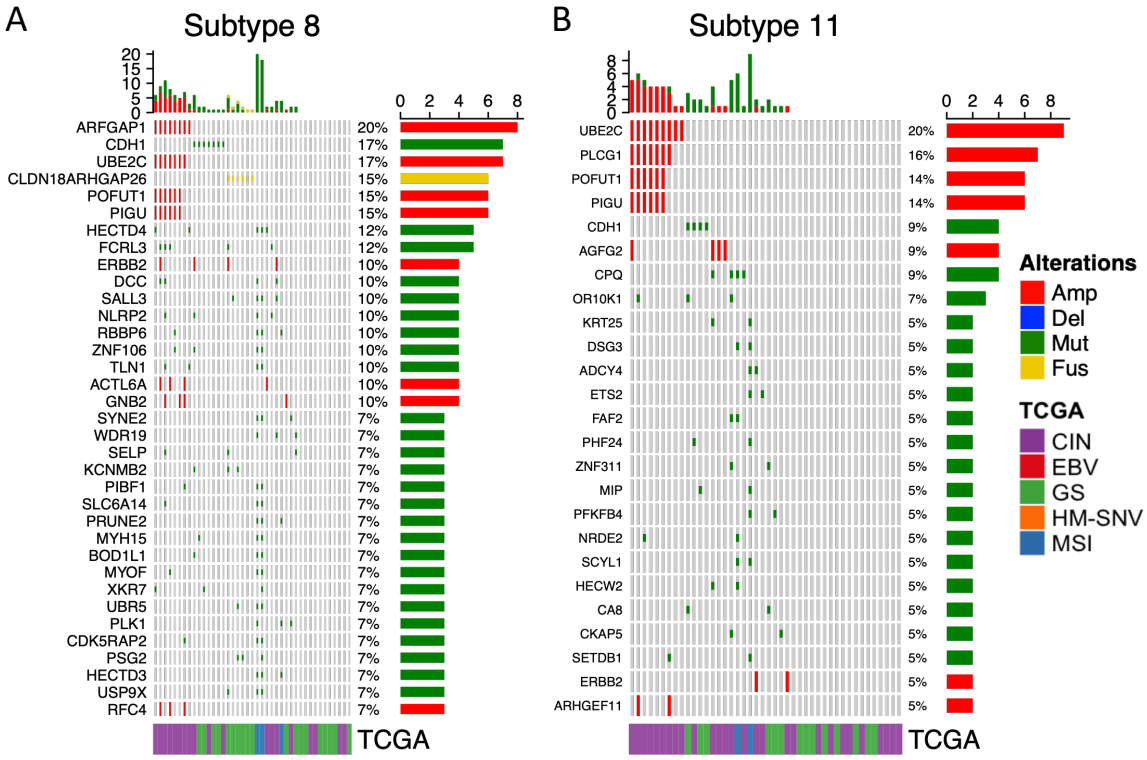


Figure 3.12 OncoPrint plots for GS subtypes.
OncoPrint plots showing predicted driver events per sample upstream of the cMRs for **(A)** S8 and **(B)** S11. Horizontal histograms and percent numbers show the fraction of samples harboring the specific event type. Vertical histograms show the number of events detected in each sample.

More interesting than the upstream events, given their sparsity, were the cMRs and their downstream targets. Gene set enrichment analysis of the top most highly predicted targets ($p < 0.05$) for S8 cMRs showed a striking pattern of immune enrichment including Reactome pathways Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell, Costimulation by the CD28 family, TCR signaling, Signaling by interleukins, and PD-1 signaling **(Figure 3.13)**.

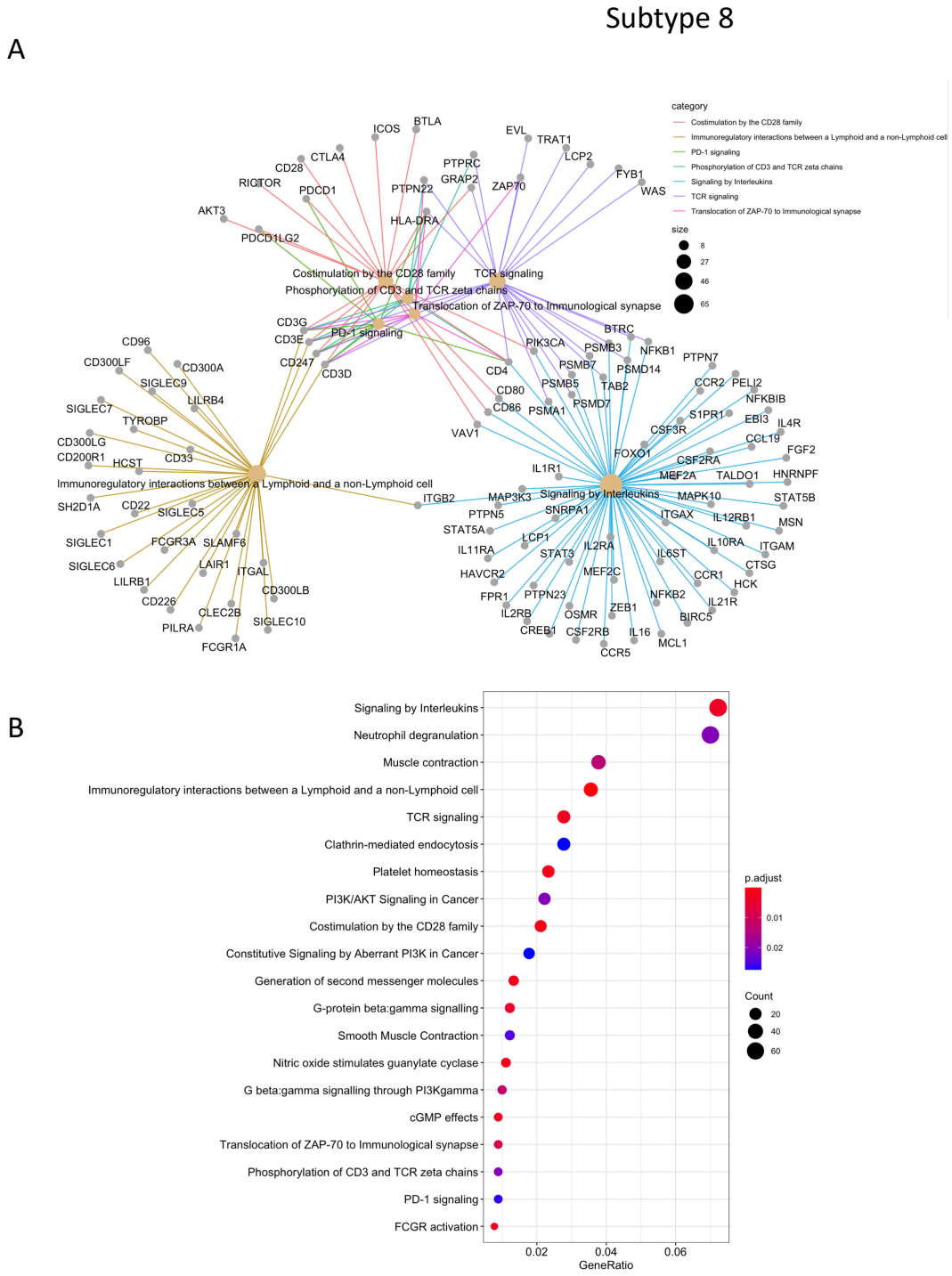


Figure 3.13 Target genes of S8 cMRs are enriched in immune pathways.
(A) Network plot of genes in immune related pathways downstream of S8 cMRs. Nodes are the Reactome pathways as labelled. (B) Top 20 most significant Reactome pathways. See Figure 3.9 for full plot type description.

A number of these targets are under the control of several immune related cMRs, specifically, FOXP3, a driver of regulatory T-cells, SPI1, a transcription involved in macrophage and lymphoid development, CD86 which is involved in T-cell co-activation and stimulation, and IKZF1 another regulator involved in lymphoid development. The enrichment of these regulators is strongly suggestive of FOXP3+ T-reg infiltration in these samples, a previously identified phenotype in a subset of gastroesophageal tumors[197]. T-regs are known to suppress other effector immune cells and can thus facilitate an immune evasive tumor microenvironment. This is further verified by the fact that this cohort of tumors has a higher percentage of leukocytes but a lower infiltration of M1 macrophages as compared to the rest of the samples ($p = 3.5x10^{-8}$ and $p = 6.7x10^{-5}$ respectively by Student's T-test) (**Figure 3.14**). T-regs have also been implicated as part of the pathogenesis of gastric cancer in patients infected with H. pylori that leads to persistent gastritis. Unfortunately for this cohort at the time of analysis too few patients were sampled for H. pylori to assess for enrichment in this subgroup. Another notable feature of this subtype is that the samples were less stem-like and more differentiated as compared to the rest of the cohort (**Figure 3.14**) ($p = 9.0x10^{-20}$, by Student's T-test)[135,189] .

Comparatively in S11, the top downstream targets of its cMRs were predominately enriched for pathways related to cellular junctions including, Extracellular matrix organization and proteoglycans, Muscle contraction, Collagen formation and Integrin cell surface interactions. Subsetting to look just at the cMRs that were predicted to be dependencies in cell lines most similar to this subtype (ZFP36L1, SCX, TCEAL3, EMX2, SMARCD3 and TCEAL7) reveals an even further refinement of similar pathways in the corresponding targets including Muscle contraction, Neuronal system, Nitric oxide stimulates guanylate cyclase and Smooth muscle contraction (**Figure 3.15**). Notably none of these cMRs have been associated with muscle or neuro-muscular

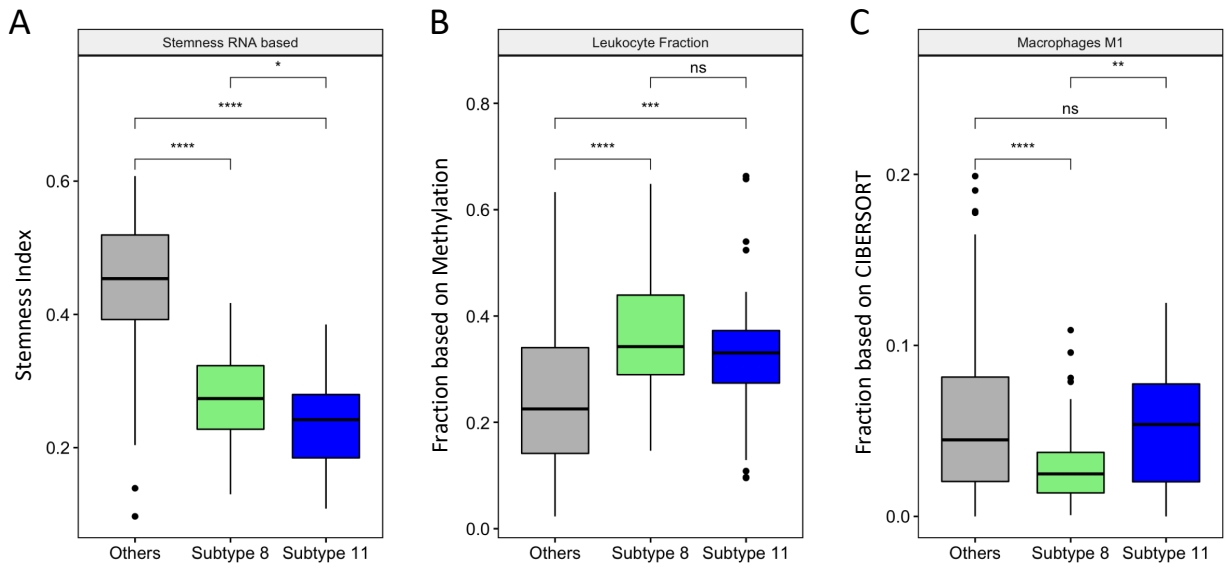


Figure 3.14 Phenotypic features of GS subtypes.
(A) Box plots of leukocyte fraction between each subtype vs all others. (B) Box plots of relative stemness. **(C)** Box plots of M1 Macrophages. All as reported in [189]. ns: $p > 0.05$, *: $p <= 0.05$, **: $p <= 0.01$, ***: $p <= 0.001$, ****: $p <= 0.0001$.

pathways before but this suggests that dysregulation of these MRs potentially co-opts these pathways. This is also further evidence that this subtype is predominantly differentiated tumors, and analysis of purity of these samples did not indicate that they were significantly less pure than the rest of the cohort. Similar to S8 this subtype had increased leukocytes and decreased stemness as compared to the rest of the cohort ($p = 3.2x10^{-3}$ and $p < 1x10^{-16}$ respectively by Kolmogorov–Smirnov test) but did not have significantly fewer macrophages or other immune cell types (**Figure 3.14**).

An analysis of the VIPER scores for both subtypes' cMRs showed that while the exact cMRs for each subtype are different, their collective activities across both subtypes are high (**Figure 3.11**). This could indicate that these subtypes have broadly similar biology but that their few driver events lead to different key cMRs.

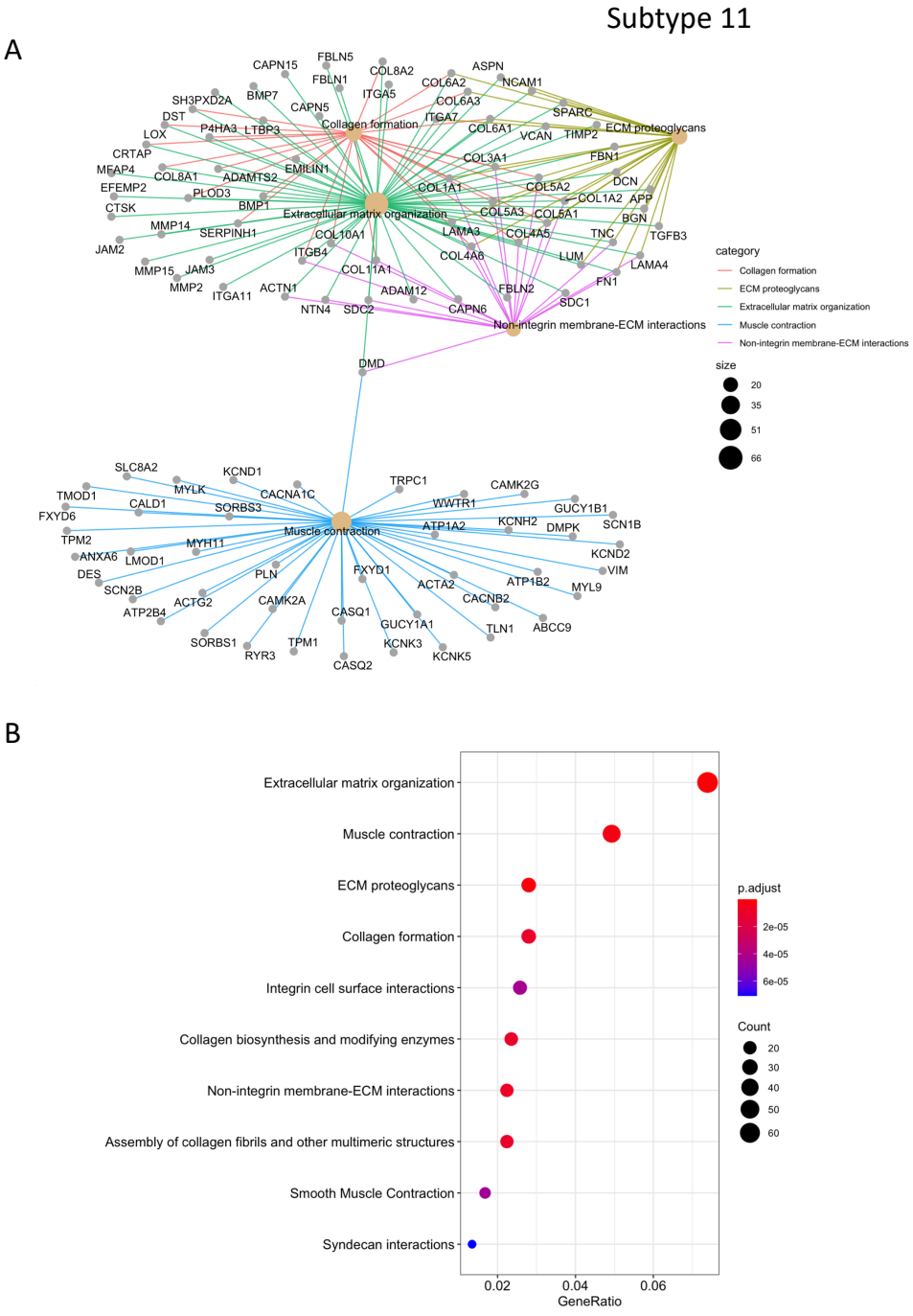


Figure 3.15 Target genes of S11 cMRs are enriched in cellular junction pathways.
(A) Network plot of genes in cellular junction pathways downstream of S11 cMRs. Nodes are the Reactome pathways as labelled. **(B)** Top 10 most significant Reactome pathways. See Figure 3.9 for full plot type description.

3.2.5 HER2+ Subtypes: S1 and S2

The other phenotype of interest was HER2 amplification. As mentioned previously, HER2 positivity is the only current biomarker used to select for specialized treatment for gastric cancer patients, but its overall efficacy is variable and limited. To try and better understand the heterogeneity present in HER2+ samples I did further analysis of the two subtypes statistically enriched for those samples, S1 and S2 ($p = 0.00012$ and 0.048 respectively by Kolmogorov–Smirnov test on GISTIC HER2 amplification scores). In addition to being enriched in HER2+ samples, S1 and S2 were also enriched in CIN type samples ($p = 2.7x10^{-8}$ and 0.009 by Fisher's exact test). Comparing the subtypes to one another did not reveal significant difference in survival, nor was there a difference between the samples that were HER2+ in each subtype versus those that were not (**Figure 3.16A-C**).

Saturation for these subtypes occurred at 46% and 42% (**Figure 3.16)**. Both subtypes were dominated by amplifications and deletions, as is to be expected because of the abundance of CIN samples, but the MOMA analysis was able to prioritize genes that are more likely to be the key drivers within a large region. As was done in the first MOMA analysis, copy number variant alterations were first filtered for whether or not they were "functional," as in whether or not the variant corresponded to an actual difference in expression. Then during the saturation calculation, proximal events within the same cytoband sub-region were combined in order to avoid double counting genomic alterations that were likely part of the same large event. Finally, to select genes most likely to be the true drivers, the gene within each region that both had the highest DIGGIT aQTL score with the cMRs identified for that subtype and was also predicted to be a CINDY modulator was determined most likely to be the true driver[83]. Interestingly for S1 of the 32 samples that did not have HER2 amplification, all but 6 had a genomic event in another gene in the HER2

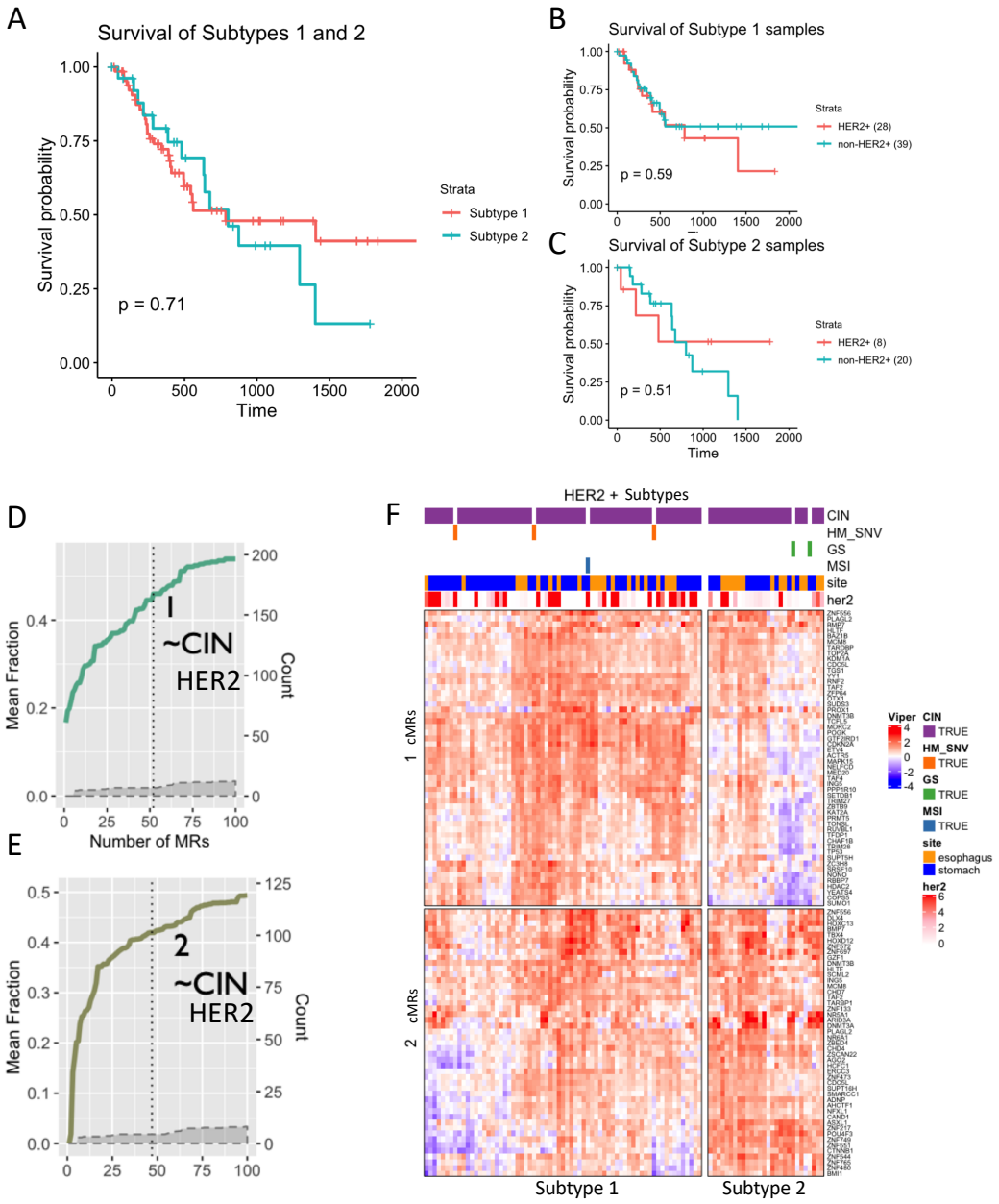


Figure 3.16 Summary of HER2+ Subtypes.
(A) Survival of 2 HER2+ subtypes. **(B-C)** Survival of HER2+ samples across all both subtypes vs the non-HER2+ samples. **(D-E)** Saturation curves as described in Figure 3.4. **(F)** Heatmap of VIPER activities of the checkpoint MRs as described in Figure 3.6.

pathway (as identified by the Reactome "Signaling by ERBB2(HER2)" gene set). As can be seen at the top of the genomics plot for S1, samples without HER2 amplification had combinations of

amplifications in PLCG1, SRC, AKT2 and AKT1, as well as a few with deletions in RHOA (**Figure 3.17A**). This is good evidence that these samples without HER2 amplification are still achieving the same biological phenotype via a different genomic event in the same pathway. In S2, this is similarly the case with many of the non-HER2 samples having amplifications in PLCG1 (**Figure 3.17B**). Gene set enrichment analysis across the other genomic events showed that both S1 and S2 had significant enrichment in events related to regulation of TP53 which is aligned with the fact that 52% and 54% of samples in these subtypes respectively have TP53 mutations. Other highly enriched pathways for genomic events of S1 included SUMOylation and pathways related to apoptosis (Death Receptor Signaling, Activation of BH3-only proteins, Activation of BAD and translocation to the mitochondria) likely indicative of aberrancies related to TP53 and other genes leading to loss of effective regulation of cell death pathways. S2 did not have the same enrichments but upon closer inspection many of the samples across S2 did have some of the same genomic events driving these enrichments in S1, they were just not predicted to be drivers for S2's cMRs.

Gene set enrichment analysis of the top most significantly predicted downstream targets of the cMRs of S1 showed that the most enriched Reactome pathways were those related to cell cycle, DNA replication and repair, and metabolism of proteins and RNA. Specifically, some of the most notable were: Cell Cycle Checkpoints, DNA Replication, Extension of Telomeres, Transcriptional Regulation by TP53, Metabolism of non-coding RNA, and SUMOylation (**Figure 3.18**). The first three were somewhat predictable as they are related to tumor proliferation, and Transcriptional regulation by TP53 again aligned with the fact that TP53 is both mutated in 52% of samples and is a cMR for this subtype. The fact that SUMOylation appeared again, in addition to enrichment in the upstream targets, was interesting as this is a relatively understudied component to gastroesophageal cancer biology and provides a potentially novel mechanism to target.

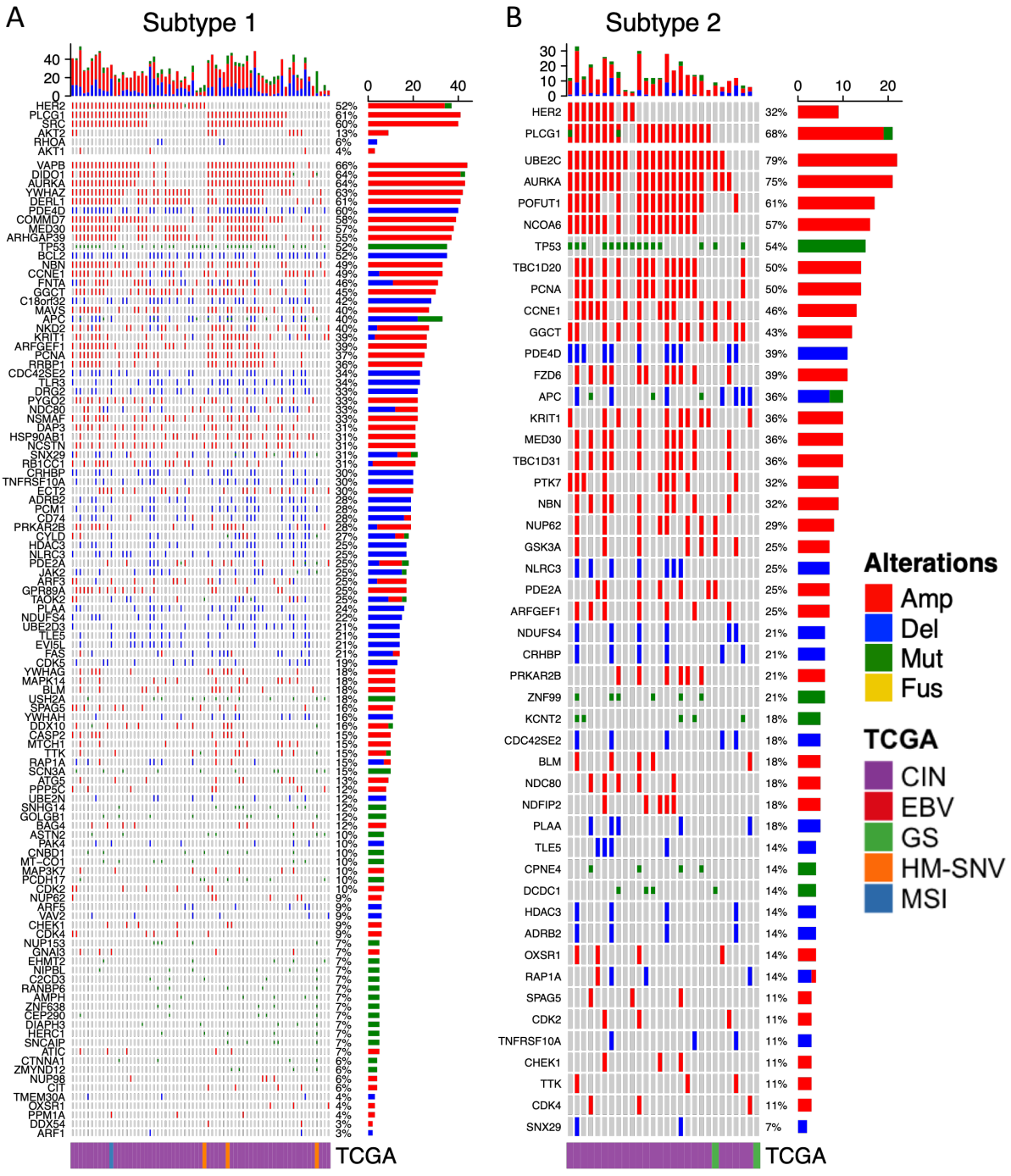


Figure 3.17 OncoPrint plots for HER2+ subtypes.
OncoPrint plots showing predicted driver events per sample upstream of the cMRs for **(A)** S1 and **(B)** S2. Horizontal histograms and percent numbers show the fraction of samples harboring the specific event type. Vertical histograms show the number of events detected in each sample. Genes in the top of both plots are part of the Reactome "Signaling by ERBB2(HER2)" gene set.

SUMOylation is a process of post-translational modification similar to ubiquitination that serves an important role in cellular response to stress and has been found to be broadly dysregulated in a number of cancers. The specific cMRs that came up in this pathway were SUMO1, one of the key components of SUMOylation activity in the cell, as well as CDKN2A, DNMT3B, RNF2, HDAC2, TRIM28, TP53, TRIM27, and TOP2A. Though its role in cancer is not fully elucidated, SUMOylation has been implicated in a number of different cancer pathways including genotoxic stress, inflammatory signaling, hypoxia, and pluripotency acquisition[198]. A broader analysis of the top 200 most dysregulated TRs overall also showed enrichment for immune related pathways, Signaling by Interleukins and Interferon gamma signaling. These enrichments were driven by down regulated TRs which could correspond with the fact that these samples had significantly fewer leukocytes and M1 macrophages as compared to the whole cohort ($p = 3.9x10^{-14}$ and $p = 6.0x10^{-3}$ respectively by Student's T-test) (**Figure 3.19**).

Interestingly, one of the most significant pathways from the same gene set enrichment analysis applied to the downstream targets of S2 was also SUMOylation (**Figure 3.20**). A post-hoc analysis showed that this similarity in enrichment is largely due to any entirely different set of cMRs and their corresponding targets. The S2 cMRs associated with SUMOylation are NR5A1, DNMT3A, BMI1, and DNMT3B, the last of which is the only one to occur in both checkpoints. This suggests a possible convergence being achieved by different sets of cMRs and transcriptional processes, and moreover that this is occurring across both HER2 amplified samples and transcriptionally similar tumors as well. Other top Reactome gene sets were M Phase, DNA Repair, Export of Viral Ribonucleoproteins from Nucleus and Antiviral mechanism by IFN-stimulated genes.

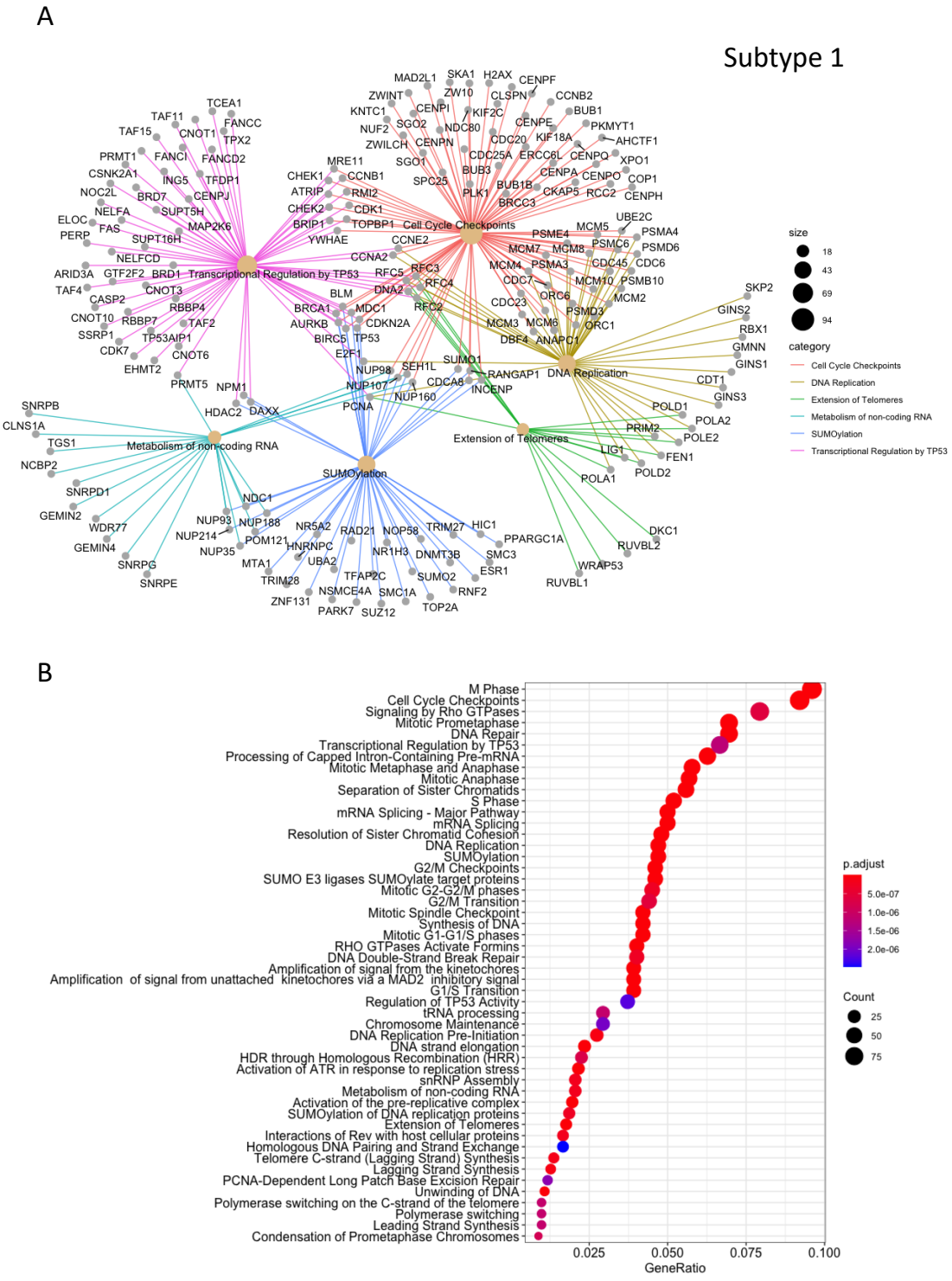


Figure 3.18 Target genes of S1 cMRs pathway enrichment.
(A) Network plot of genes downstream of S1 cMRs. Nodes are the Reactome pathways as labelled. **(B)** Top 50 most significant Reactome pathways. See Figure 3.9 for full plot type description.

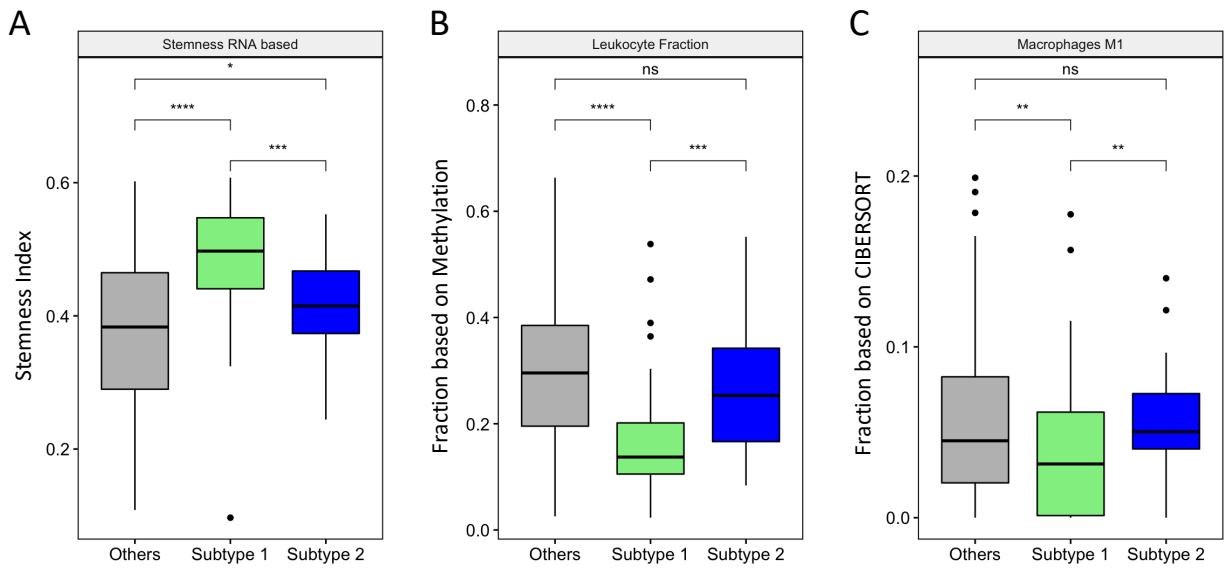


Figure 3.19 Phenotypic features of HER2+ subtypes.
(A) Violin plots of leukocyte fraction between each subtype vs all others. **(B)** Violin plots of relative stemness. **(C)** Violin plots of M1 Macrophages. (D) Violin plots of resting NK cells. All as reported in [189]. ns: p > 0.05, *: p <= 0.05 , **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001.

3.2.5.1 HER2 & Drug Resistance

Trastuzumab is a monoclonal antibody that targets HER2 and is the only currently approved biomarker specific drug for gastric cancer patients, but it is minimally effective and acquired resistance happens frequently. To address these issues, other HER2 targeting therapeutics have been investigated as potential alternatives. Lapatinib, a tyrosine kinase inhibitor, was shown to be effective in trastuzumab-resistant breast cancer but a clinical trial in gastric cancer patients did not show the same efficacy[199–201]. Afatinib, an irreversible TKI that binds to EGFR, HER2 and HER4, was approved as a therapy for non-small cell lung cancer and early studies in HER2+ gastric cancer cell lines and PDX models indicate that afatinib has better anti-tumor and anti-metastatic activity as compared to lapatinib[199,202,203].

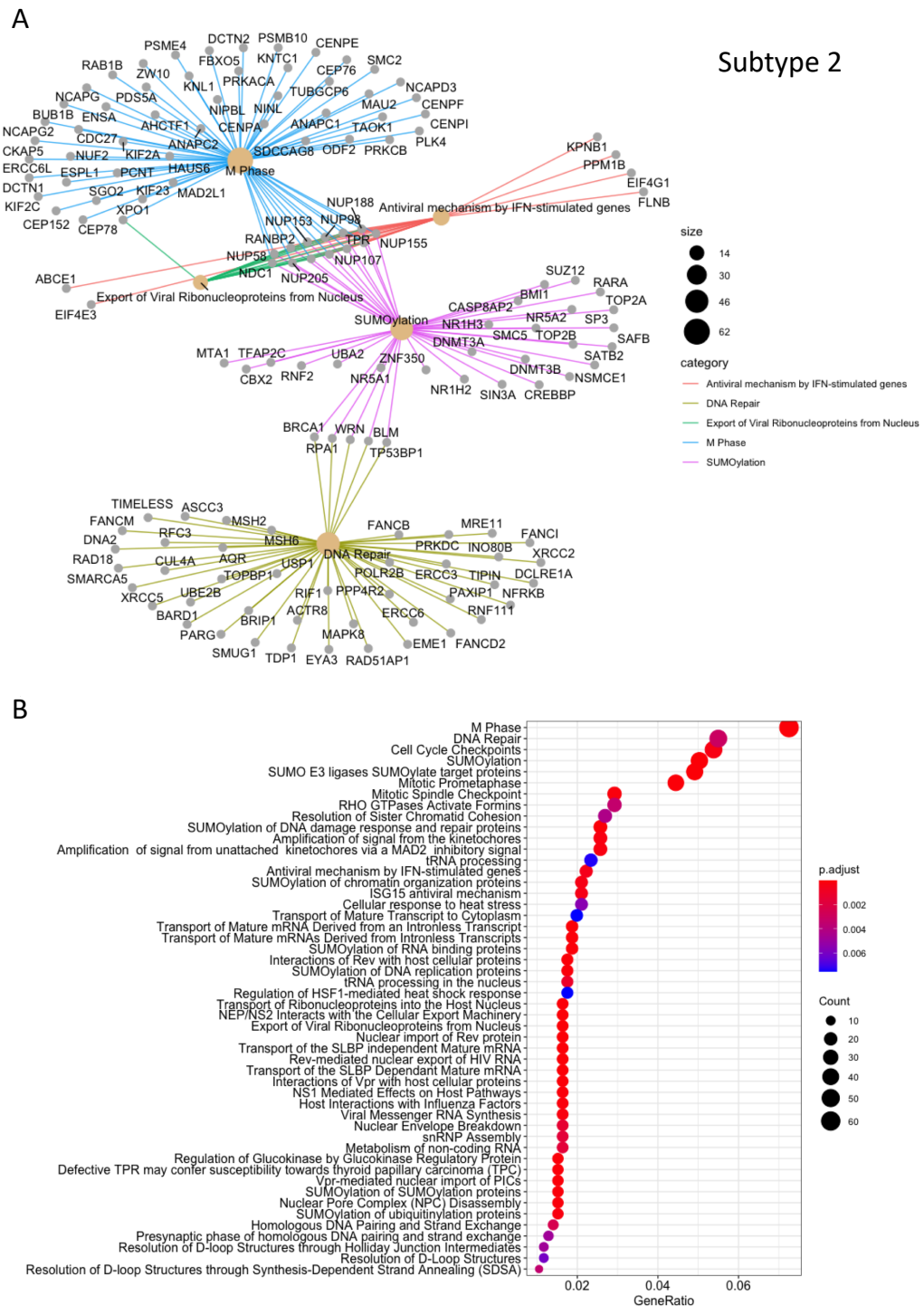


Figure 3.20 Target genes of S2 cMRs pathway enrichment.
(A) Network plot of genes downstream of S2 cMRs. Nodes are the Reactome pathways as labelled. **(B)** Top 50 most significant Reactome pathways. See Figure 3.9 for full plot type description.

Work done by our collaborators in the Bass laboratory has been ongoing to better characterize the mechanisms and therapeutic potential of afatinib in HER2+ gastroesophageal cancer samples. Interestingly, in their studies they have found that while afatinib is effective in two HER2+ cell lines, NCI-N87 and OE19, its effect is markedly decreased when these cell lines are grown in three-dimensional scaffolding (Alvetex). Notably, this is not the case in a HER2+ breast cancer cell line, BT474. Joint delivery with belinostat, an HDAC inhibitor, is able to rescue the effect of afatinib, through it is not effective when used alone, implying some level of synergy at play. As based on these observations and to test some of my hypotheses from the MOMA pipeline, I performed MR analysis on samples from a drug screen across these three cell lines.

3.2.5.2 Experimental Design and Brief Summary of Results

The three cell lines of interest, OE19, NCI-N87 and BT474, were grown in both 2D and 3D cell culture. Alvetex was used for the scaffolding structure in the 3D experiments, which allowed for the same growth media conditions to be used for both. The three drug conditions tested were afatinib alone, belinostat alone and combined afatinib and belinostat. DMSO, the vehicle for delivery of these drugs, was used for the control condition. The cells were treated for both 6 hours and 48 hours in order to assess both short- and long-term effects of the drugs. All conditions were done in triplicate. **Table 3.3** summarizes the drug conditions and efficacy.

Table 3.3 Experimental Design and Results of HER2+ Drug Screen

Dimension	Drug	BT474	OE19	N87
2D	Afatinib	Yes	Yes	Yes
	Belinostat	--------	Yes	Yes
	Combined	--------	Yes	Yes
3D	Afatinib	Yes	***No ~60% cells remaining***	***No ~20% cells remaining***
	Belinostat	-------	***No***	***No***
	Combined	-------	Yes	Yes

3.2.5.3 Using an MR-Based Classifier to Predict Drivers of Drug Resistance

Raw fastq files were acquired after the samples were sequenced by the Broad Institute. Read counts were then calculated using Kallisto and transformed using variance stabilizing transformation from the DESeq2 package (refs). I performed an initial principal component analysis in order to confirm that all triplicates were reliably similar. In doing so I found two outlier samples (a BT474 3D sample treated with Afatinib for 6 hours and an OE19 3D sample treated with Belinostat for 48 hours). These were removed from further downstream analyses.

In order to generate VIPER predicted protein activity scores, I paired each experimental condition sample with the corresponding set of DMSO controls in that cell line at that time point in the same condition, ie to generate a signature for OE19 3D samples treated with Afatinib for 6 hours, I used OE19 3D samples treated with DMSO for 6 hours. This was done to adjust for any contributions that may have arisen because of the conditions themselves and to optimize detection of drug specific MRs. After generating these signatures, I then performed single sample VIPER

analysis on each experimental sample using the STES interactome from the MOMA analyses for the gastric cell lines and an interactome built on TCGA BRCA samples for the BT474 samples. Hierarchical clustering inclusive of all tested TRs showed that cell line and time were still the most significant drivers of similarity, not drug sensitivity (**Figure 3.21**).

Based on this observation, I decided to build a multi-step classifier to determine which TRs were predicted to be key regulators of drug sensitivity across multiple conditions. To do this I selected two sets of condition pairs in the gastric cancer lines that were most informative for interrogating the regulators of the sensitivity: 1) 2D afatinib treatment (sensitive) vs 3D afatinib treatment (resistant) and 2) a combination of all sensitive conditions (2D afatinib, 2D afatinib with belinostat and 3D afatinib with belinostat) vs 3D afatinib (resistant). For each cell line I performed

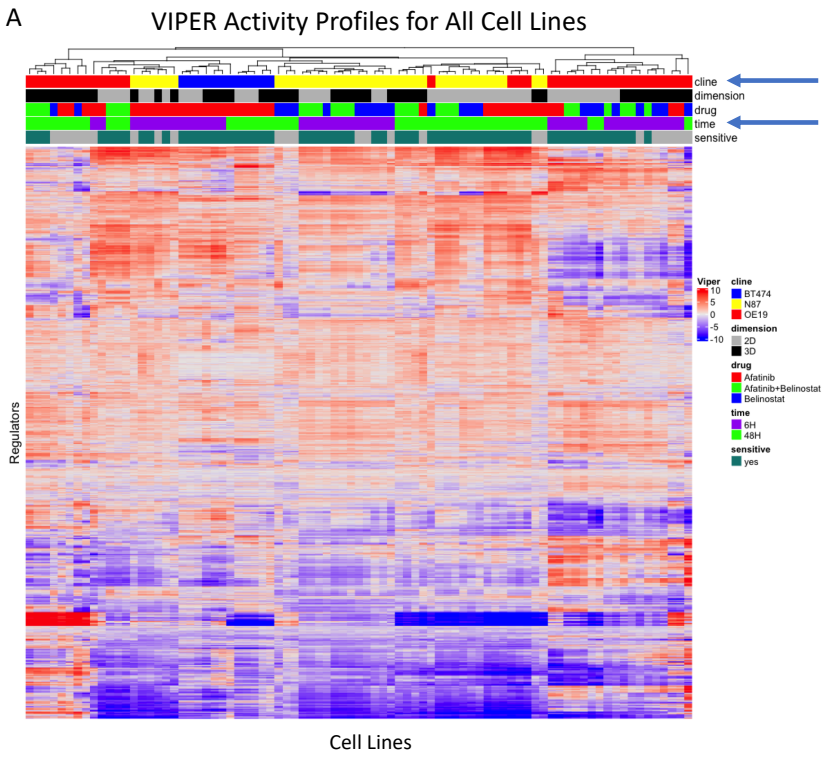


Figure 3.21 Heatmap of VIPER activity across all cell lines show that dominant drivers of difference are cell line and time.
Hierarchical clustering of resulting VIPER profiles for each cell line. Rows are TRs and columns are cell lines. Top annotation describes attributes of each cell line. See legend for colors.

linear discriminant analysis across each TR to determine which ones were able to discriminate between the two conditions with 100% accuracy. Examination of the overlap between both cell lines and across both condition pairs revealed 12 TRs. Filtering these 12 TRs down to the ones that shifted in the same direction activity-wise between sensitive and resistant, resulted in 9 candidate MRs ($p = 0.0001$ after fitting a null distribution to 10^6 random TRs selected to as classifiers after shuffling sample labels). See **Figure 3.22** for the number of TRs identified at each step. For this analysis I only considered samples from the 6-hour incubation time point as principal component analysis showed that all the 48-hour time point samples had a very strong separate signal of activity seemingly unrelated to drug their drug resistance status. I also reasoned that the 6-hour time point was a better reflection of the immediate transcriptional response to treatment of the drugs.

The 9 final MRs fell into two groups in terms of biological mechanism: histone 3 lysine (H3K) methylation regulation (ASH2L, SNAI2, and JARID2) and cell differentiation and proliferation (CDCA7, MYB, WNT5, ESX1, KLF7 and HOXA10). ASH2L is a methyltransferase that methylates H3K4 depending on its H3K9 methylation status. It's been shown to interact with the TAF family of proteins as well as MYC in certain contexts. SNAI2 has been shown to work with KDM1A, a histone demethylase, to decrease dimethylated H3K4 and repress transcription, in addition to being involved in the induction of epithelial to mesenchymal transition. JARID2 is a repressor that recruits the PRC2/EZH2 histone methyltransferase complex and contributes to gene regulation in embryonic stem cells. In terms of the other group of MRs, MYB (an oncogene), CDCA7, and KLF7 are all involved in pathways regulating cell proliferation and differentiation. WNT5A is involved in regulating the Wnt signaling cascade, depending on the context, and HOXA10 is known to interact with PTPN6, EGFR and STAT3/STAT1 during oncogenic transformation processes. ESX1 seems to be involved in spermatogenesis but also is broadly

associated with the GATA family of genes. Finding both of these signals, methylation regulation and cell proliferation is consistent with the fact that belinostat is an HDAC inhibitor, which has

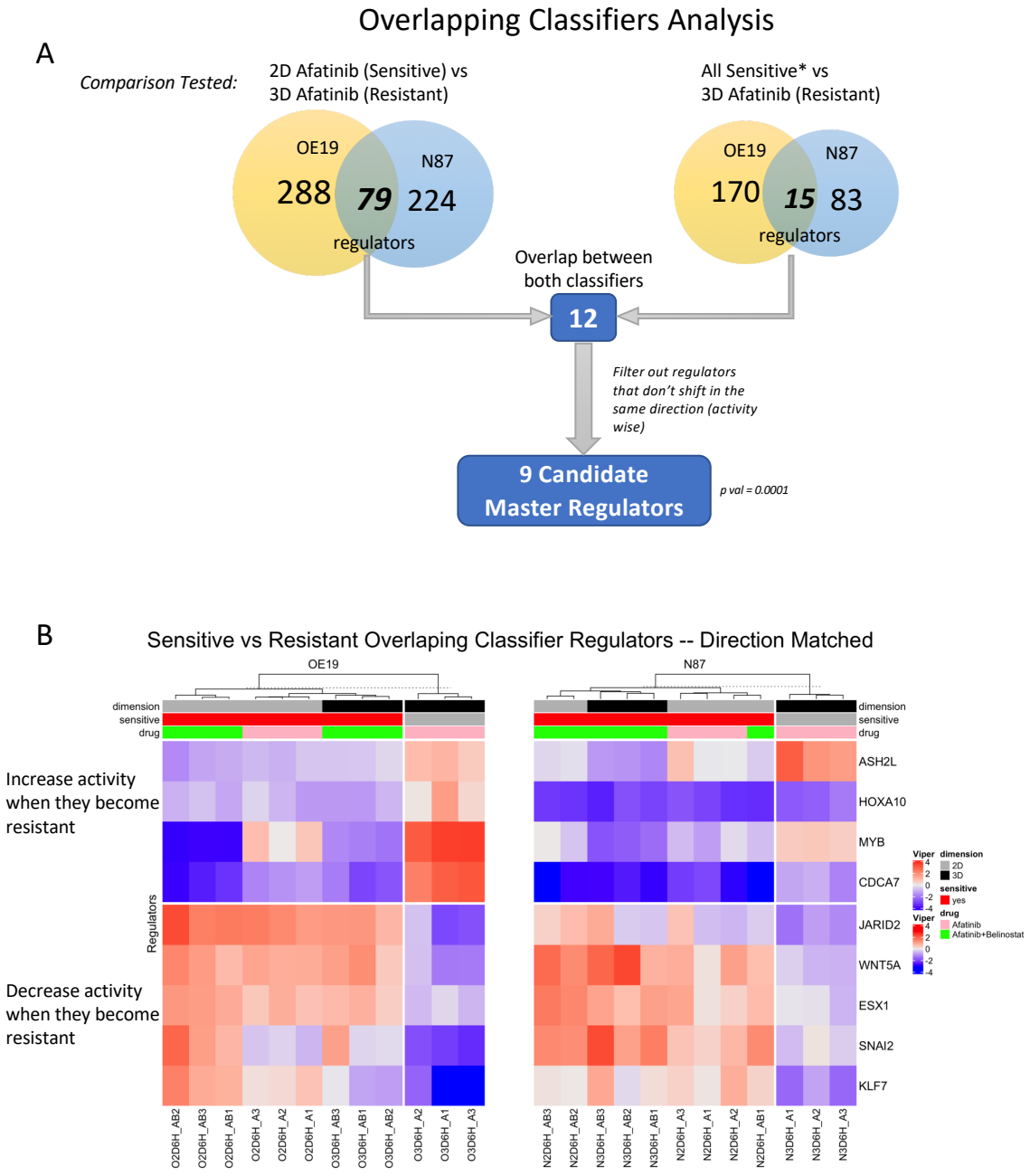


Figure 3.22 Schematic of multi-step classifier and resulting top TRs.
(A) Venn diagram of TRs selected as classifiers per cell line per comparison. **(B)** Heatmap of resulting top classifier TRs split by cell line.

been shown to have effects across multiple types of histone modifications and chromatin structure, and because the drugs are aimed at blocking proliferation of the cancer cells. Work is ongoing to understand more about the role of these 9 candidate MRs, as will be discussed later.

3.2.5.4 Using Condition Paired VIPER Analysis to Predict Drivers of Drug Resistance

In order to cast a wider discovery net to understand the biology underpinning this resistance, and to select TRs to test in a corresponding CRISPR screen, I performed a second set of analyses on this data using a different methodology to prioritize candidate MRs. Rather than using a classifier to identify individual TRs capable of differentiating between drug resistant and sensitive, I instead set up several paired VIPER analyses that allowed me isolate and interrogate the drug specific phenotypes. More specifically, I first generated a gene signature between samples reflecting the change in drug sensitivity, ie between samples treated with afatinib in 3D (resistant) vs 2D (sensitive), then used VIPER to generate a ranked list of regulators. I then also generated a gene signature of the "background," meaning the aspect of the previous signature not related to the drug activity, but rather related to the cell type and conditions. For example, for the afatinib in 3D vs 2D VIPER analysis, I generated a background signature and VIPER profile using samples with only DMSO delivered in 3D vs 2D. I then took the ranked list of TRs and their relative activity scores from the main condition and subtracted the activity scores inferred from the background condition to generate an adjusted set of VIPER scores for each TR now having accounted for biological changes not related to the drug sensitivity (Figure ***). In addition to the above condition pair I also generated adjusted VIPER profiles for the combination drug condition (afatinib and belinostat) in 3D culture with adjusted backgrounds for both afatinib and belinostat alone. Separately paired analyses were done for each cell line and time point and top TRs that were

statistically significant across multiple conditions were selected as candidates ($p < 0.05$ after FDR correction).

Gene set enrichment analyses of the candidate MRs that came up in at least one third of the pairs showed that the top most enriched Reactome pathways were Chromatin modifying enzymes and organization, Transcriptional regulation by TP53, HATs acetylate histones and Regulation of lipid metabolism by peroxisome proliferator-activated receptor alpha (PPARalpha). The high enrichment of chromatin modification related TRs is consistent with the fact that belinostat is an HDAC inhibitor and aligns with the predictions from my first classifier, where 3/9

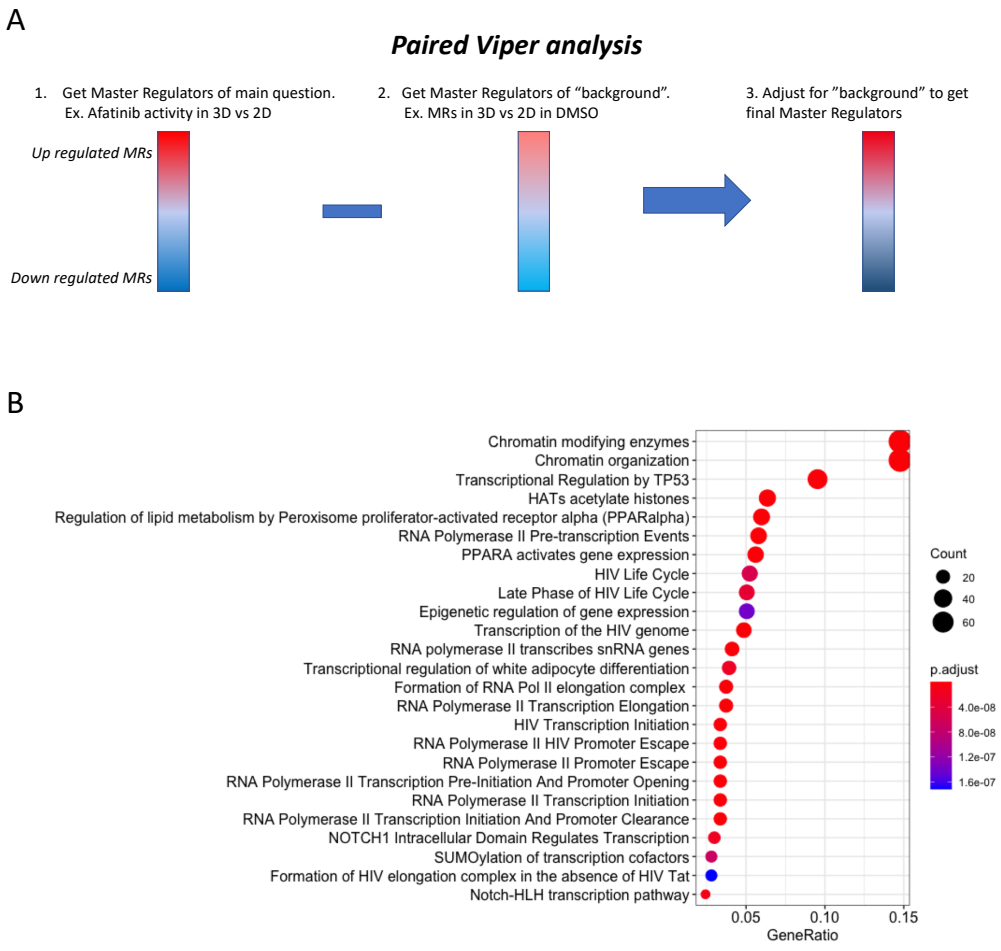


Figure 3.23 Paired VIPER Analysis to select candidate MRs.
(A) Schematic of work flow. **(B)** Gene set enrichment of TRs that came up as significant in at least 1/3 of the conditions.

predicted MRs were related to histone methylation. Notably SUMOylation of transcription cofactors also came up in the top 25 most enriched pathways in part because of SUMO1 and BMI1, which I also identified as cMRs in my MOMA analysis. This gives further credence to the potential importance of these TRs and the role of SUMOylation in this condition. To further validate these predictions all TRs that were significant in at least 2 or more conditions were selected to be tested in a follow up CRISPR screen. That screen is currently underway so the results were not available at this time.

3.2.6 Cell Line Matching to MOMA Inferred Subtypes

To further analyze the biological importance of the cMRs, I looked for enrichment of cMRs as relatively essential genes from the Achilles analysis in cell lines that were high quality matches for the different subtypes of interest. To test for patient similarity, all cell lines from the CCLE that were annotated as being derived from either the esophagus, stomach or upper aerodigestive tract were selected as candidates (101 cell lines in total). An internal gene signature was generated across this pool of samples in order to amply the differences among them and to mirror the internal signature used for the MOMA analysis. VIPER analysis was then applied to transform the signature into inferred protein activity scores for each cell line using the STES interactome.

Two methods were used to assess similarity between patients and cell lines. In one, I matched each patient to each cell line in a pairwise manner using viperSimilarity to test for enrichment of the top and bottom 25 MRs, as was done in the previously described Achilles analysis. Cell lines were considered matches if the enrichment was significant at a threshold of $p < 0.01$ after multiple hypothesis correction. The second method used to assess similarity was to first create a single representative profile for each subtype before doing the similarity enrichment analysis. To do this I first applied Stouffer's Integration across the TRs of the samples in each

subtype, then subsequently used this integrated profile to generate a subtype match score with each cell line as previously described. For this analysis I focused on the cell lines that optimally matched S1 and S8 as they had a higher enrichment of HER2 and GS samples respectively.

These two methodologies produced mostly congruent results (**Figure 3.24**). Many of the top most predicted cell lines for S1 using the two different methods were derived from esophageal squamous cell carcinomas. This made them less desirable based on their cellular origin but does open up the possibility of biological similarity at the transcriptional level. The 7th best match using the second method was SNU16 a cell line derived from a gastric carcinoma which also broadly matched a number of patients in subtypes 4, 5 and 6 in addition to S1. GSEA analysis of S1 cMRs on the Achilles CRISPR based essentiality scores for SNU16 showed significant enrichment of these cMRs ($p = 0.010$), further confirming the importance of these as key regulators. The 2nd best match for S8 using both methodologies was HS746T, a cell line derived from a gastric adenocarcinoma. Repeating the same analysis using the S8 cMRs also showed significant enrichment of these MRs as being likely essential in HS746T ($p = 0.021$) (**Figure 3.25**).

Using these matching methodologies also revealed that while NCI-N87 and OE19 were good matches for HER2+ patients, they did not match the non-HER2+ patients in S1 and S2 particularly well. This suggests that there may be some limits to the broader usability of some of the insights from the drug screen analyses previously mentioned. Additionally, a second CRISPR screen for top MRs in a best matching GS matching cell line are also in preparation.

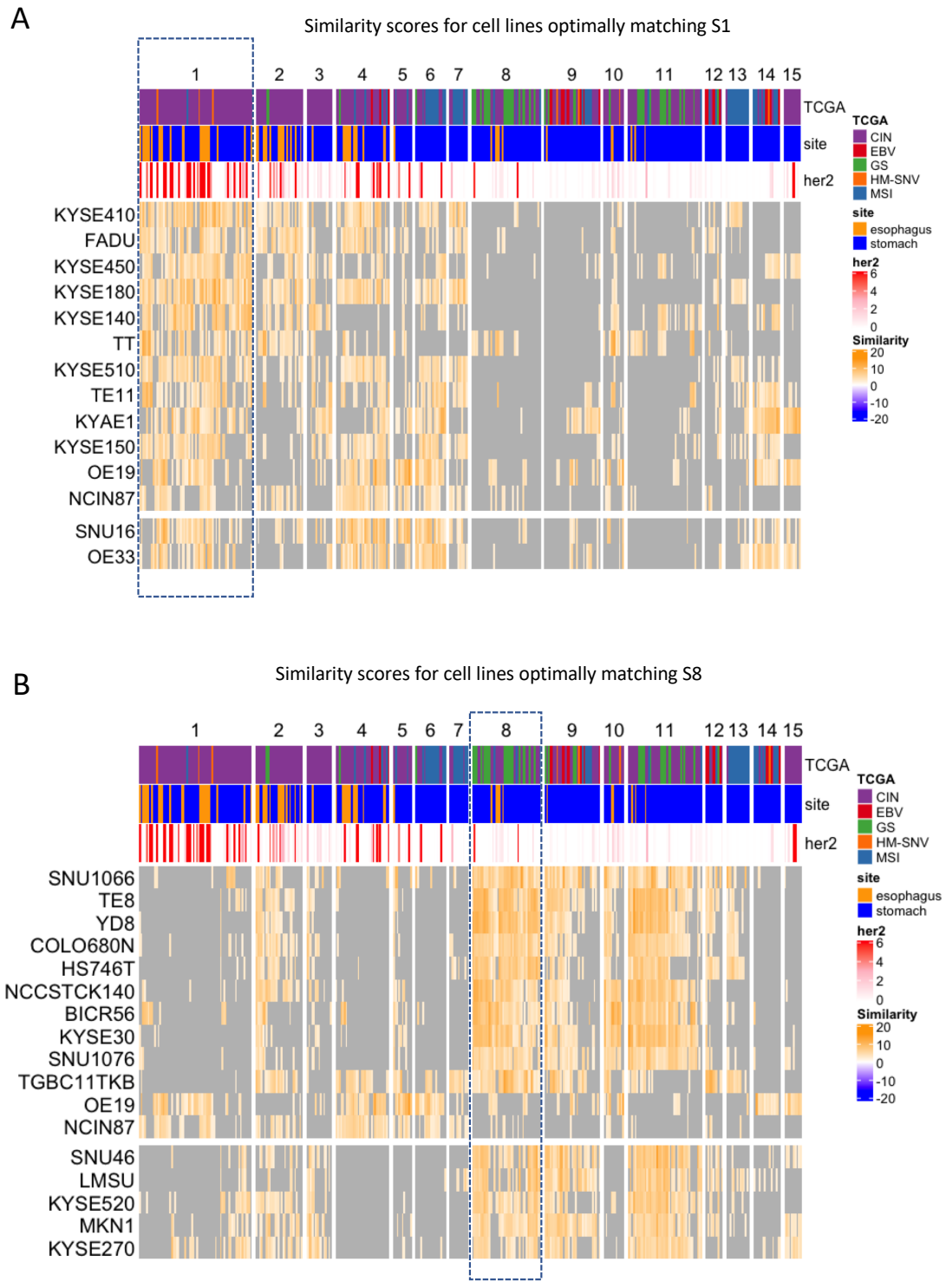


Figure 3.24 Top matching cluster specific cell lines.
(A) S1 and **(B)** S8**.** Patients are columns and cell lines are rows. For clarity only significant matches are colored ($p < 0.01$ after multiple hypothesis correction). Top annotations are the same as previous plots.

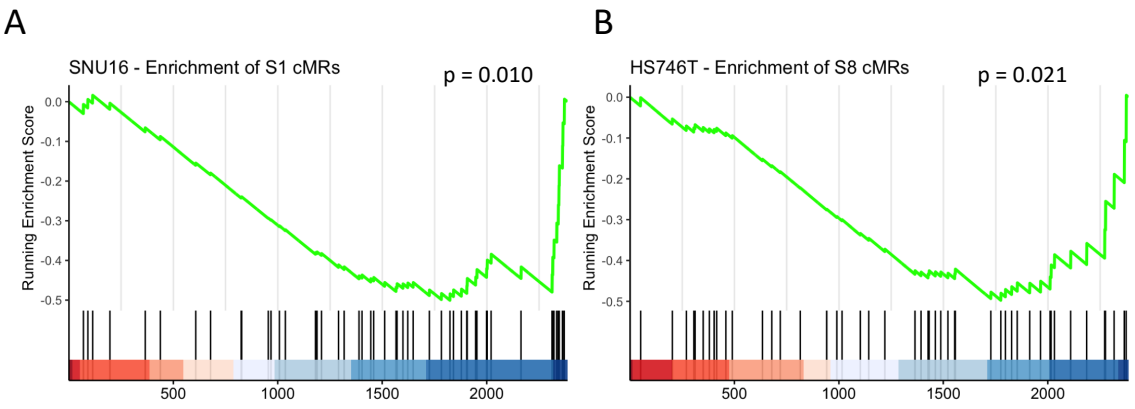


Figure 3.25 GSEA Enrichment of cMRs in Achilles scores for matching cell lines.
(A) Enrichment of S1 cMRs on ranked Achilles scores for SNU16 (negative indicating a gene is more likely to be essential). **(B)** Enrichment of S8 cMRs on HS746T Achilles scores.

3.2.7 Validation in External Cohort

With any classification system built on a single dataset, the risk of overfitting specifically to that data is high. In order to determine whether or not my new methodology and system was more generally applicable to gastroesophageal adenocarcinoma biology, I used the data from the Asian Cancer Research Group (ACRG) as a validation set. As mentioned previously in the introduction, this research group has also developed their own gastric cancer classification system as based on a cohort of 300 patient samples acquired at the Samsung Medical Center[192–194]. Their classifier is based on gene expression and delineated four subtypes: microsatellite unstable, mesenchymal-like (EMT), TP53 active, and TP53 inactive. The first two subtypes approximately align to the MSI and GS subtypes, respectively, from the TCGA analysis (which they confirmed in their work) but the TP53 based subtypes were less closely aligned with any particular subtype.

To begin this analysis, microarray data from the GEO database (GSE62254) was downloaded and transformed to VIPER inferred protein activity values using the STES interactome. To match the STES analysis the differential signature was generated using an internal reference (comparing the samples to one another) so as to highlight the intra-cohort differences.

Principal component analysis of the resulting ACRG VIPER profiles along with the STES samples showed that the VIPER transformation was able to almost entirely mitigate batch effects (**Figure 3.26**). Applying the Global MOMA weights across the TRs of each of the cohorts (as was done prior to the clustering step in the STES analysis) resulted in even better concordance between the two.

In order to understand how well their classification labels corresponded to the subtypes identified by the MOMA analysis, I built a random forest classifier to predict the different ACRG labels using VIPER protein activity values as features. After selecting 75% of the samples for the

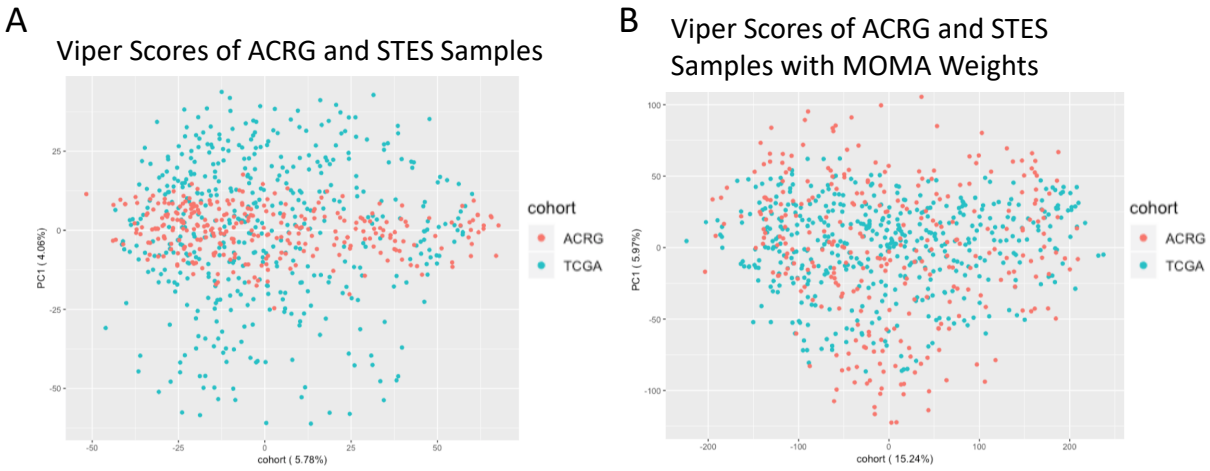


Figure 3.26 PCA plots of patient's VIPER profiles from each cohort.
(A) Unadjusted VIPER profiles used for each patient. **(B)** VIPER profiles after weighting with STES global MOMA scores.

training, I did fivefold cross validation for ten iterations, and then applied the resulting model to the test group. When considering all the regulators as features I found that the overall accuracy ranged from 45.5% to 87.5%, with a median of 66.7%. Closer inspection revealed that the EMT and microsatellite unstable labels were being accurately predicted close to 100% of the time but the accuracies for the TP53 active and inactive labels were quite low. To try and improve the overall model accuracy I utilized the Boruta algorithm to prioritize selection of the most

statistically significant relevant features[204]. Boruta is an all-relevant feature selection method that generates a set of randomly permuted shadow features and includes them in the training step of the random forest. These serve as a null to compare to the actual features and from this it is possible to delineate which features performed statistically better than the randomly permuted ones. Using this method refined my list of features down to 121 TRs from the original pool of 2445. I then built another random forest model only including these TRs as features and this resulted in accuracies ranging from 55.7% to 79.2%, with a median accuracy of 68.2%. Attempts at using other classifier methods did not lead to improvements in accuracy so I selected the random forest model based on

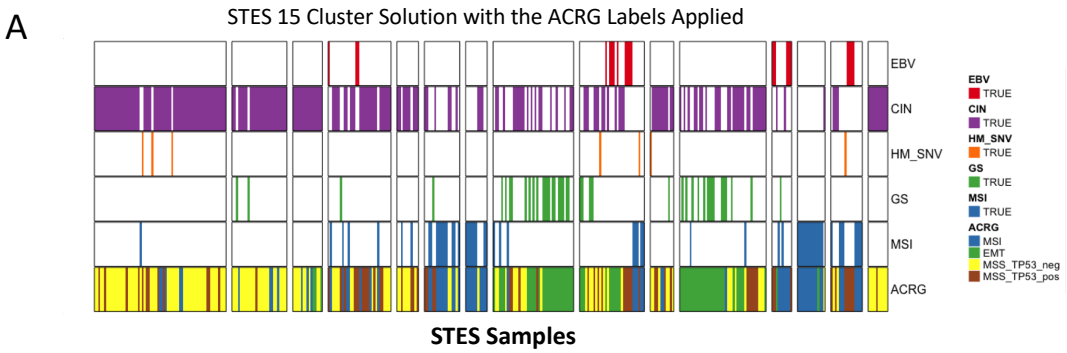


Figure 3.27 Subtype annotation plot of STES samples with ACRG labels applied.

the Boruta selected features. Applying this model to the STES samples resulted in concordance of the two expected pairs of subtypes, EMT with GS in S8 and S11, and microsatellite unstable with MSI in S6, S7 and S13 (**Figure 3.27**).

To further understand whether or not the MOMA and iterative clustering methodology could be validated in this cohort I applied it to the VIPER profiles of the ACRG cohort. Using the same global MOMA weights from the STES analysis to first weigh the TR features in the ACRG cohort I then proceeded with iterative clustering using the same parameters as described previously. The ACRG samples stabilized at 11 clusters with EMT samples almost entirely

represented in one cluster and the microsatellite unstable distributed across three clusters. All of these clusters had high activity of the expected cMRs as identified in the MOMA analysis, and overall had the same global patterns of cMR activity (**Figure 3.28**). Though fewer overall subtypes were found by the iterative clustering method this is likely due to overall differences in the cohort make up, and suggests that the TCGA cohort has a wider breadth of patient types.

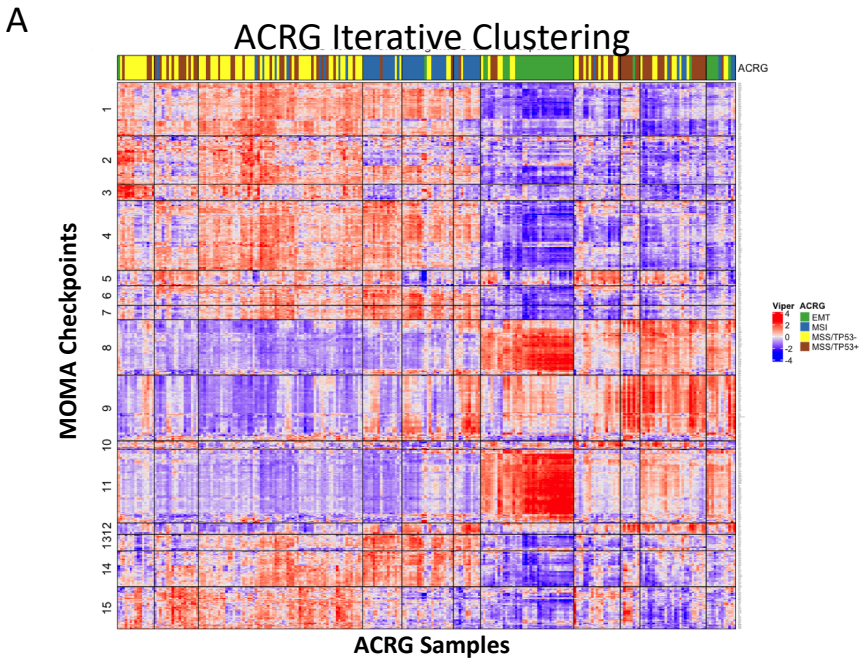


Figure 3.28 Results of iterative clustering applied to ACRG Samples.
Row are the MOMA cMRS inferred from the STES cohort. Columns are ACRG patients. Top annotations are the ARCG patient labels.

3.2.8 MOMA Identification of Global Regulators of Gastroesophageal Cancer

Though most of the analyses I performed utilized VIPER inferred values from a gene signature generated internal to the STES cohort, this minimized the ability to find likely global drivers of the tumors as it was designed to accentuate intracohort differences. In order to interrogate the broader biology present across these samples I generated a gene signature using all of the tumor samples in the TCGA as a reference group, similar to what was done for each cohort

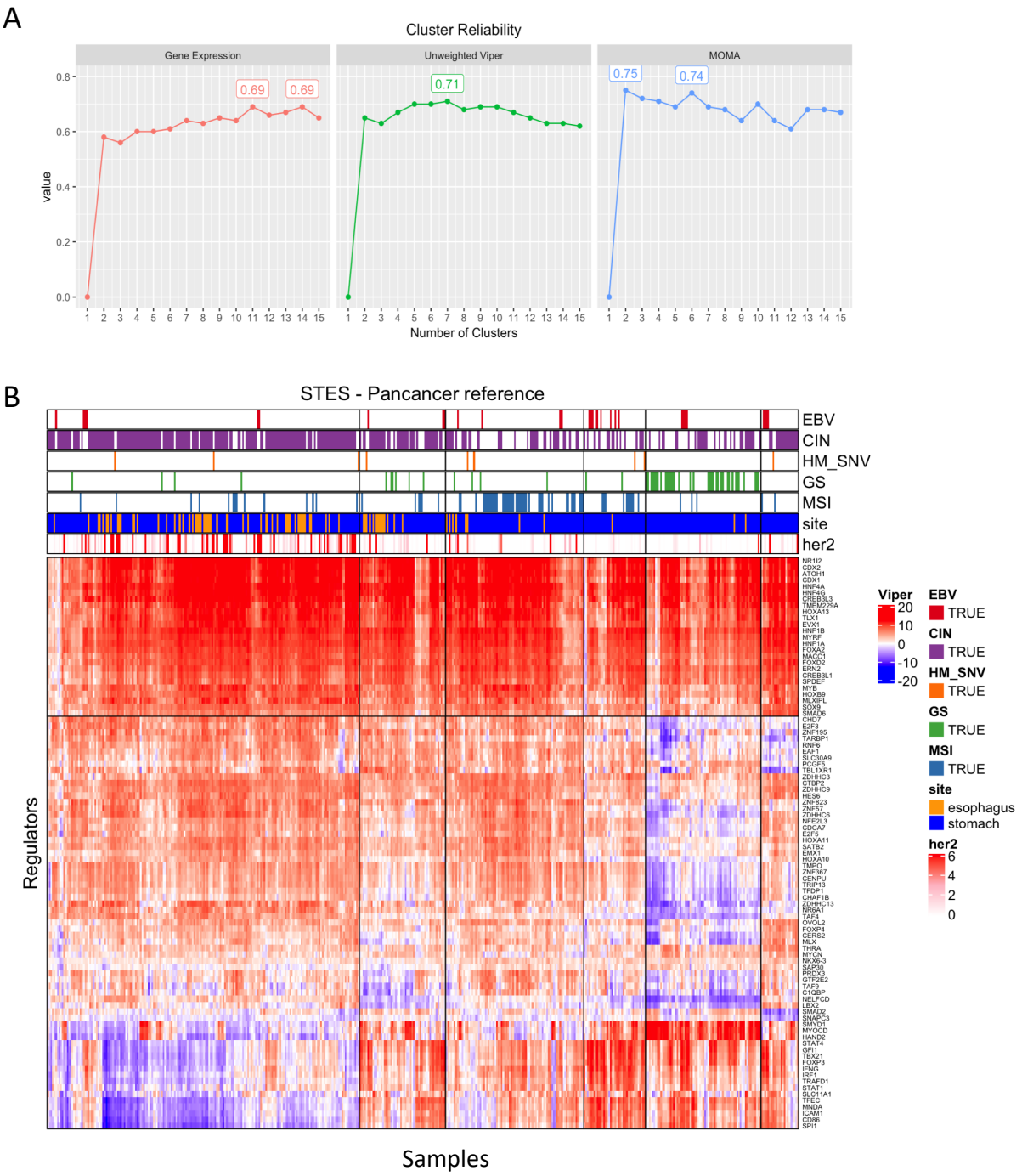


Figure 3.29 Clustering Results using TCGA reference VIPER activities.
(A) Clustering results comparing average cluster reliability scores for each clustering solution using k = 2 to k = 15. Panels show clustering using gene expression, unweighted viper scores and MOMA weighted VIPER scores respectively. **(B)** Heatmap of 6 cluster solution based on TCGA reference. Top annotations as previously described.

in the original MOMA analysis. In doing so I was able to capture not only the differences among the samples but also which TRs are likely to be key MRs for gastroesophageal cancer as a whole.

After generating this TCGA-reference signature I again transformed the values to generate VIPER inferred protein activities for all of the TRs using the STES interactome. MOMA analysis was then performed as previously described in order to generate a ranked list of MOMA scores for each TR and then I proceeded with clustering using these weights. A comparison of cluster reliability scores (see previous chapter's methods for description of this metric) showed improved average cluster similarity scores when using VIPER as compared to gene expression and were even better when using MOMA weighted VIPER values (**Figure 3.29**). This analysis showed that 2 and 6 clusters were statistically equivalent so I selected the 6-cluster solution as it showed substantial enrichment of the 3 subtypes of interest, HER2+, MSI and GS, in clusters 1, 2 and 5 respectively (see **Table 4**).

Table 4: Enrichments of TCGA subtypes in 6 cluster solution using TCGA reference.

Subtype	# of Samples	Final Average Cluster Reliability Score	Enrichments					
			CIN	HM-SNV	MSI	GS	EBV	HER2
1	192	0.891	1.6e-16	0.96	1	1	1	1.0e-09
2	53	0.506	0.69	0.96	1	1	1	1
3	85	0.542	1	0.96	1.1e-13	1	1	1
4	38	0.598	1	0.96	0.21	1	0.0016	1
5	71	0.936	1	1	1	5e-20	1	1
6	23	0.412	0.76	0.96	1	1	0.16	1

Analysis of the resulting cMRs across these 6 subtypes revealed that approximately a quarter of them (25/90) had uniformly high predicted activity across all of the patients. Of these 25, 5 of them are known to be associated with gastrointestinal development (HNF4G, HNF1B, HOXA13, FOXA2 and ERN2) and 9 have been identified as gastroesophageal cancer biomarkers

(HNF4A, CDX2, HNF1B, HOXA13, FOXA2, MACC1, MYB, HOXB9, and SOX9). Identification of these known associated regulators further confirms that this analysis is identifying true biological signals. Additionally, it suggests that these and the other cMRs identified as being highly dysregulated across these samples could be potential therapeutic targets.

3.2.9 Using Precision Oncology Algorithms to Predict Novel Therapeutics

In order to prioritize drugs that target these global dependencies, I first performed cell line matching analyses as described previously. VIPER activity signature for gastroesophageal cell lines from the CCLE were computed from gene signatures that compared them to all other available cell lines. This was done to optimize comparability to the TCGA reference STES profiles as described above. Matches between a patient and a cell line were determined at a threshold of $p < 10^{-10}$ after Bonferroni correction, **Table 5** shows the resulting best cell lines by overall percentage match to patients, and **Figure 3.30** shows the top 6 based on per cluster matching percentage. NCI-N87 was selected for screening as it matched a substantial number of patients (78.1%) based on this analysis, and it would be possible to couple it with the other analyses done to investigate HER2 amplification.

Table 5: Best cell line matches across the STES cohort.

Rank	Cell Line	Patient Coverage
1	X2313287	85.1%
2	TGBC11TKB	83.1%
3	KATOIII	81.8%
4	NCIN87	78.1%
5	KE39	77.5%
6	NUGC4	77.3%
7	AGS	76.0%
8	GSU	75.8%
9	SNU719	75.8%
10	MKN45	72.7%

Using the precision oncology pipeline OncoTreat developed in the Califano lab, NCI-N87 cells were perturbed with 336 different FDA Phase I/II approved compounds at sublethal doses for

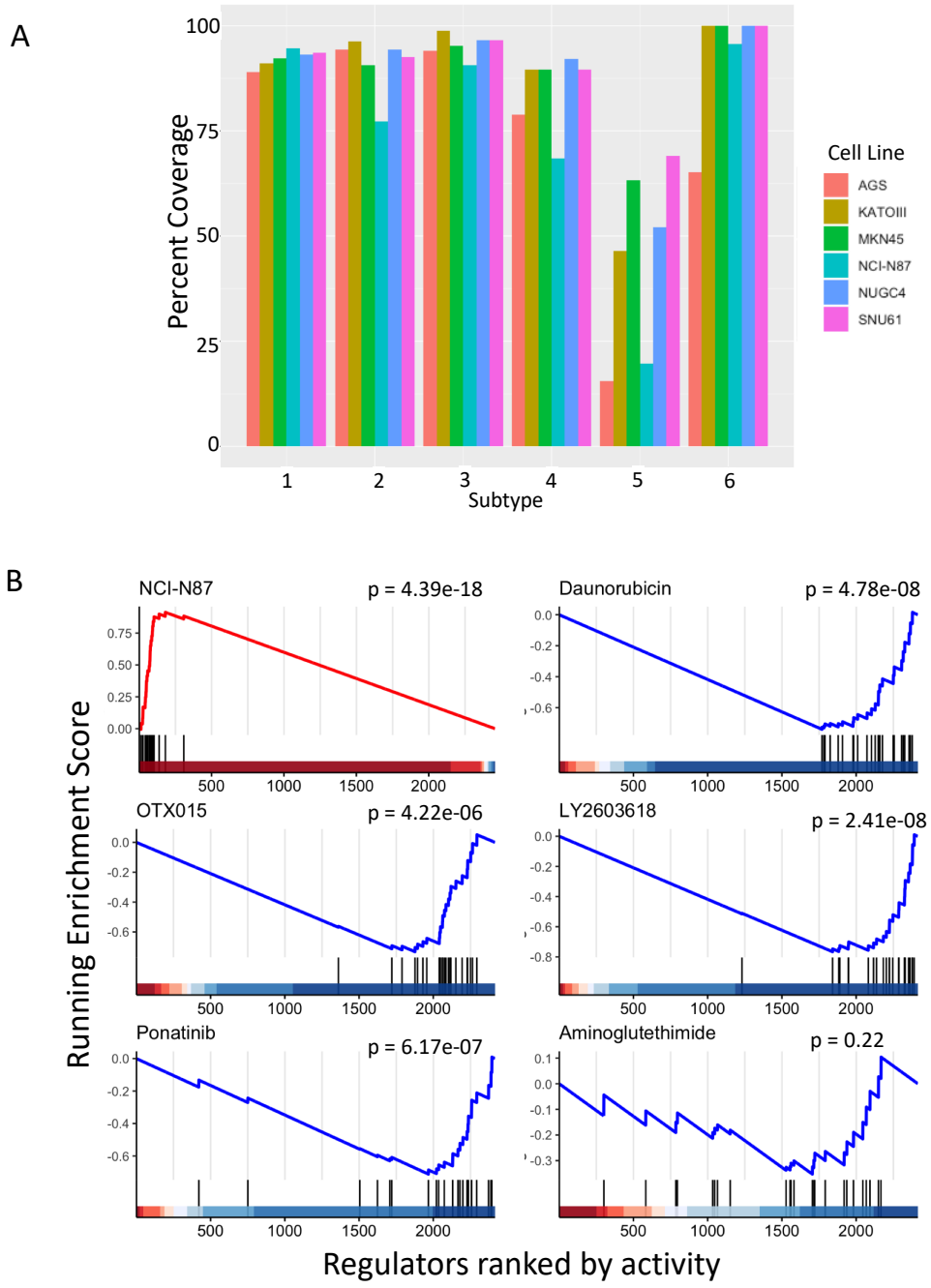


Figure 3.30 Prioritizing drugs based on global STES cMRs.
(A) Percent coverage of matching cell lines for each subtype. **(B)** OncoTreat results of 6 highly predicted drugs for inverting NCI-N87's VIPER signature. GSEA based on top 25 cMRs for STES.

24 hours and then subsequently sequenced[130]. The resulting RNAseq profiles were then transformed to VIPER activity scores for each drug condition, reflecting the regulatory changes that occurred in response to exposure to the drug. With this information I then determined which drugs most significantly inverted NCI-N87's VIPER activity profile by calculating the enrichment of its top 25 positive and 25 negative regulators that were then inactivated and activated by each compound treatment, respectively. A select set of the top 20 drugs predicted by this analysis are plotted in **Figure 3.30** along with the ranks of the top 25 MRs that were identified as being highly dysregulated across all patients. A number of the top most drugs – Daunorubicin, and its analog Idarubicin, Teniposide, Irinotecan and Doxorubicin – are known to interfere with topoisomerase activity and Irinotecan in particular is already part of treatment regimens for patients with metastatic colorectal carcinomas. Daunorubicin has been plotted as a representative drug for this class. Other top predicted drugs included OTX015, a bromodomain inhibitor, LY2603618, a selective CHK1 inhibitor, and Ponatinib, a multi-target kinase inhibitor. Aminoglutethimide, an aromatase inhibitor sometimes used to treat breast cancer, was at the top of the list in terms of inverting NCI-N87's VIPER signature, but was the worst of these in terms of specifically inverting the 25 MRs of interest. This highlights the value of looking not only at the effect on the cell lines but also prioritizing for inversion of the key MRs of interest. Notably none of these drugs are currently used for treatment of gastric cancer so could all be further pursued as novel therapeutic options. Work is also underway to screen a second cell line that matches S8 patients (GS enriched subtype) in order to prioritize drugs for those patients as these were not well represented by NCI-N87.

3.3 Discussion

Gastroesophageal cancer is one of the most commonly diagnosed malignancies worldwide. Despite this much is still unknown about the etiology and development of these heterogenous tumors, thus stymying development of precise and effective therapies. Over the years several groups have developed genetics-based classifications for identifying different subtypes of these tumors but they have not yet led to effective changes in the clinical landscape.

This work utilized the Master Regulator framework as developed in the Califano lab in order to identify and interrogate the key transcriptional dependencies driving different subtypes of gastroesophageal cancer. Building upon the multi-omic methodology as elucidated in the first MOMA paper, I sought to characterize not just the master regulators themselves but also to identify the key genomic drivers leading to their aberrant activity. In doing so, I aimed to capture a fuller picture of the complex biology driving the oncotecture of these tumors, which has been yet to be comprehensively done.

Using a refined version of the MOMA framework along with an iterative clustering methodology identified 15 different subtypes. Though generated in an unsupervised manner these subtypes aligned with previously described biological phenotypes, giving credence to this method's ability to both capture established patterns as well as discover new ones. Analyses of checkpoint MRs, i.e. those able to canalize a high proportion of the likely driver genes across a particular subtype, showed a statistically significant enrichment for relatively essential genes, suggesting that these cMRs do represent key potentially targetable "Achilles heels" of these tumor subtypes.

Exploration of the resulting subtypes revealed complex webs of biology at play. In the case of the MSI dominant subtypes, S6, S7, and S13, driver mutations in genes related to extracellular

matrix organization and connectivity, along with a higher degree of predicted stemness suggest that the mutations in these tumors facilitated progression through epithelial-mesenchymal transition. The cMRs of these subtypes were overall very similar and controlled cell pathways related to cell replication and DNA synthesis, as well as pathways implicated in response to HIV infection, possibly due to the abundance of neo-antigens in these hypermutated tumors.

The GS enriched subtypes, S8 and S11, exhibited minimal driver mutations, as was to be expected. Key genomic events in S8, including ARFGAP1 (amplification), CDH1 (mutation) and CLDN18-ARHGAP26 (fusion), occurred in a largely orthogonal manner indicating that a number of these patients had aberrancies in different pathways but still converged to similar transcriptional profiles. This is a key benefit of the MOMA framework as it is harder to identify these similarities when looking only on a gene by gene basis, particularly when very few mutations are present. S8 had a strong signal of immune infiltration, opening up the possibility of effectively targeting these patient types with combination immune checkpoint inhibitor therapies. Part of the signal seemed to be driven by T-regs which may complicate this therapeutic strategy as they have been implicated in acquired resistance to these therapies, so a more complex regimen may be required (Saleh 2019). S11 on the other hand had a higher percent infiltration of leukocytes but did not have pronounced dysregulation of T-reg cMRs, suggesting they may be more susceptible to standard immune checkpoint inhibitor therapies.

Analysis of the two HER2+ enriched subtypes revealed that many of the samples without HER2 amplifications instead harbored mutations elsewhere in the HER2 pathway, likely leading to their transcriptional similarity. Because of the high frequency of chromosomal aberrancies across these samples, the DIGGIT/CINDY component of the MOMA framework was particularly crucial for prioritizing candidate drivers within these regions. Enrichment of genes in the

SUMOylation pathway occurred across the pool of upstream genomic drivers as well as in the set of genes under the control of the cMRs, suggesting increased dependence on this intracellular signaling mechanism. Further analysis of drug resistance to Trastuzumab in HER2 cell lines suggested that this process may be driven by chromatin and DNA modifying regulators, particularly histone demethylases, but the results of a follow up CRISPR screen for validation are not yet complete. Preliminary results from a different screen done in the lab, in which a panel of key TRs were knocked out and subsequently sequenced to interrogate the expression of their targets, suggests that a number of the key cMRs for S1 may coordinate each other's activities as a hyperconnected module.

Combining pan-cohort MR predictions with the OncoTreat drug prediction algorithm, identified a number of novel drug classes that significantly inverted activity of these key regulators, providing clinically pursuable avenues for treatment of these tumors. A follow up screen in a second cell line as well as in patient derived organoids is underway in order to further refine and prioritize these candidate therapies.

In addition to corresponding with subtypes identified by our collaborators at the Broad institute, the subtypes predicted by the MOMA analysis also aligned well with the biological classifications identified by ACRG[189,194]. Validating my predictions in this external dataset provides strong evidence that the biological patterns MOMA identified are not purely a result of over fitting to the primary dataset. Though this is promising, further work still needs to be done to build true a MR based classifier on the STES subtypes. This will be crucial for both further validating its applicability to other datasets as well as confirming its ability to be used as a tool for new patient classification. Additionally, repeating this analysis on the NUS cohort will serve as further validation of this new classification system.

While this analysis provided a number of insights into the regulatory logic of gastroesophageal cancers, this study also had a number of limitations. Currently the multi-omic framework of the MOMA analysis only incorporates mutations, copy number variants and fusions as candidate genomic driver events, and in doing so misses out on other potential driving disrupters of the regulatory logic, particularly epigenetic modifications. Efforts were made to make the algorithm more flexible to additional omic information but they were not successful at improving the identification of key biological patterns. Further work will need to be done to fully achieve this goal.

Additionally, though incorporating the iterative clustering methodology revealed interesting and biologically meaningful subtypes, in certain cases it seems to have over stratified the samples resulting in subtypes that are more similar than they are different. This seemed to be the case for the MSI enriched subtypes suggesting that for clinical classification it may be more useful to recombine them into one group. That said, a number of non-MSI samples were categorized as having similar regulatory patterns, something that would have been missed if the initial analysis had not been agnostic to these labels. Moreover, while a handful of these subtypes were selected for deeper analysis a more comprehensive examination of all the subtypes may reveal even more actionable insights as well as other instances of over stratification.

As mentioned previously, a striking part of this analysis was its ability to recapture many previously identified biological phenomena in a completely unbiased manner. More follow up experimental work is underway and will hopefully validate the more novel predictions.

Discussion

4.1 General Conclusions

The advent of modern sequencing technologies undoubtedly revolutionized the field of cancer genomics and brought upon the early ages of the precision medicine era. Despite early successes with a few biomarkers, translating the biology of many tumor types into actionable and effective therapies has remained elusive. One of the issues stalling progress has been the focus on the oncogene addiction theory of cancer, a model that while true in some cases of tumorigenesis has not turned out to be the prevailing mechanism for a large percentage of tumors.

Models from the world of network biology have helped to reframe the lens by which the field interrogated the complexity of cancer. Much work has been done in an attempt to reverse engineer the dynamic, multi-layered array of interactions driving cancer cell biology. Instead of looking for singular drivers, modules of genes were instead evaluated for their role in driving cancer phenotypes. Through this lens, insights have been gleaned not only about the convergence of mutations in biologically similar pathways but also about the key role that regulatory factors play as nodes within these cellular networks.

Indeed as more data from sources like the TCGA and others has become available it's been found that in contrast to the vast genetic heterogeneity observed across cancers, the transcriptional states are remarkably similar and stable[43,44]. This suggests that subsets of tumors may accrue different genomic events but ultimately converge to the same dysregulated but homeostatic state.

These transcriptional states can be characterized by the key transcription factors, or Master Regulators, that coordinate the expression of the genes that are aberrantly expressed in the tumor state. In several individual contexts using this Master Regulator framework has been shown to be more effective for both classifying tumor subtypes as well as identifying the key regulatory factors

that are both sufficient and necessary for maintaining a particular cancer state. Prior to the work outlined in Chapter 2 of this thesis, a comprehensive pancancer Master Regulator analysis had not been done.

We developed and applied MOMA, as described in Chapter 2, a methodology built to effectively integrate together several types of omics data to better characterize the complex genomic landscapes driving these dysregulated transcriptional states. The broad goals of this analysis were two-fold: one, to test the Oncotecture Hypothesis and to determine whether tumor checkpoints could be identified on a sample by sample basis, which would integrate the genetic alterations in that sample to implement its transcriptional state, and two, to assess whether, even within tumor checkpoints, MR proteins may form smaller, highly recurrent modular structures. Using the MOMA framework, we identified 112 subtypes across 20 cohorts of the TCGA, nearly all of which had identifiable MR checkpoints linked to sample specific upstream genomic alterations. Subtypes across 19/20 cohorts showed significant survival separation, something that had not been previously achievable for certain cohorts using gene expression based clustering alone. Analysis of the most recurrent of these checkpoint MRs revealed a highly degree of modularity across them, thus confirming both parts of the Oncotecture Hypothesis. Further probing the biology of these MR-Blocks revealed functional alignment with a number of the classical Hallmarks of Cancer, two of which were validated experimentally.

In Chapter 3 of this thesis I further expanded on this MOMA framework by directly applying it to a cohort of gastroesophageal adenocarcinomas. Though one of the most prevalent and lethal tumor types, very few specialized treatments exist, largely due to the high heterogeneity of these tumors. Using an updated version of the MOMA framework in conjunction with iterative clustering revealed 15 different subtypes across the cohort that closely aligned with subtypes as

previously identified across the field, though they were identified in an unbiased manner. Moreover, this new framework captured the upstream variants across this tumor type much better than the first MOMA analysis, thus showing the value of applying MOMA to both large and small cohorts.

Analysis of clusters enriched in three of the major previously defined phenotypes—Microsatellite Instable (MSI), Genomically Stable (GS) and HER2 amplified—revealed a number of insights about the complex genetic architecture underpinning the transcriptional states of these tumor types. The MSI subtypes (S6, S7 and S13), as predicted, had the highest frequency of genomic events, particularly point mutations, in addition to amplifications across chromosome 8. The predicted driver genes were enriched in genes related to cell membrane structure and ECM connectivity indicating a potential EMT progression, further bolstered by 2/3 subtypes higher degree of stemness. The GS subtypes (S8 and S11) on the other hand had far fewer driver events, though most occurred in a mutually exclusive manner across the samples within each subtype, indicating that different "recipes" of genomic events led to the same transcriptional state. Notably S8 had significant leukocyte enrichment and its cMRs controlled a number of immune related pathways potentially indicating the possibility of targeting these patients with some form of immunotherapy. Across the HER2+ subtypes (S1 and S2), nearly all the samples without HER2 amplification instead had predicted driver events in other genes in the HER2 signaling pathway, specifically PLCG1, SRC, AKT2, AKT1 and RHOA. This explains the similarity in transcriptional profiles of the non-HER2 samples within these subtypes. Further work to better characterize and validate predicted MRs driving drug resistance to Trastuzumab, a HER2 monoclonal antibody-based therapy, are ongoing. In addition to these biological insights, applying the Califano lab's recently developed precision oncology pipeline, OncoTreat, to one of the representative cell lines

predicted a number of drugs that effectively inverted key global MRs and could be further pursued for clinical use.

4.2 Future Directions

The results of the MOMA framework provide a number of novel avenues to explore moving forwards. A database containing all ~ 2 million interactions between the regulators and predicted driver events has been made publicly available and can serve as a jumping off point to pursue newly predicted connections via experimental validation. This is similarly the case for the MR-Blocks as we only validated 2 of the 24. A deeper exploration of the biology of these MR-blocks, particularly their combinatorial nature within individual checkpoints could reveal not only novel biological insights but also pathways for new MR-block specific cancer therapeutics. Additionally, an R-package containing the MOMA pipeline has also been publicly available and can thus be applied any cohort for which matched expression and mutational data is available. We hope that the field will find this useful both in the context of cancer as well as other disease types.

Though MOMA provided a number of novel insights both from a pancancer lens as well as in a more fine-grain analysis on gastroesophageal tumors, there are certainly limitations and areas of improvement. As with other high-throughput methods, both experimental and computational, it is reasonable to expect that MOMA will also produce false positive and negative predictions. Further experimental validation of some of these predictions would need to be done to confirm the degree to which this is the case. Known issues with some of the high false negative rates in some of the underlying algorithms are currently being addressed by improving the statistical models to improve sensitivity.

MOMA could be further augmented to more flexibly integrate other omics into its framework. Attempts were made to do this for application to the STES cohort, but more work still needs to be done to best capture the underlying distributions of the various omics data in order to ensure that they are integrated in a biologically and statistically robust way.

With regards to the analyses of the gastroesophageal tumors, while many novel insights were predicted, the experimental validation for a number of them is still pending. Once acquired these will hopefully lead to actionable insights that can be used to help provide these patients with therapeutics more precisely tailored to their tumor type.

References

1. National Cancer Institute. The Genetics of Cancer. (2017).

2. International Agency for Research on Cancer. Global Cancer Observatory. Available at: https://gco.iarc.fr/. (Accessed: 16th May 2021)

3. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* caac.21660 (2021). doi:10.3322/caac.21660

4. National Cancer Institute. Cancer Statistics. (2020). Available at: https://www.cancer.gov/about-cancer/understanding/statistics. (Accessed: 16th May 2021)

5. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA. Cancer J. Clin.* **69**, 7–34 (2019).

6. Krzyszczyk, P. *et al.* The growing role of precision and personalized medicine for cancer treatment. *TECHNOLOGY* **06**, 79–100 (2018).

7. National Cancer Institute. Types of Cancer Treatment. Available at: https://www.cancer.gov/about-cancer/treatment/types. (Accessed: 16th May 2021)

8. Morgan, G. W., Ward, R. & Barton, M. The contribution of cytotoxic chemotherapy to 5-year survival in adult malignancies. *Clin. Oncol.* **16**, 549–560 (2004).

9. Spear, B. B., Heath-Chiozzi, M. & Huff, J. Clinical application of pharmacogenetics. *Trends in Molecular Medicine* **7**, 201–204 (2001).

10. Kisor, D. & Ehret, M. *THE PERSONALIZED MEDICINE REPORT*.

11. Nowell, P. C. The clonal evolution of tumor cell populations. *Science (80-.).* **194**, 23–28 (1976).

12. Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197–200 (1975).

13. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **340**, 1546–1558 (2013).

14. Dang, C. V. MYC on the path to cancer. *Cell* **149**, 22–35 (2012).

15. Hajdúch, M., Jančík, S., Drábek, J. & Radzioch, D. Clinical relevance of KRAS in human cancers. *Journal of Biomedicine and Biotechnology* **2010**, (2010).

16. Hantschel, O. & Superti-Furga, G. Regulation of the c-Abl and Bcr-Abl tyrosine kinases. *Nature Reviews Molecular Cell Biology* **5**, 33–44 (2004).

17. Zilfou, J. T. & Lowe, S. W. Tumor suppressive functions of p53. *Cold Spring Harbor perspectives in biology* **1**, (2009).

18. Dyson, N. J. RB1: A prototype tumor suppressor and an enigma. *Genes and Development* **30**, 1492–1502 (2016).

19. Song, M. S., Salmena, L. & Pandolfi, P. P. The functions and regulation of the PTEN tumour suppressor. *Nature Reviews Molecular Cell Biology* **13**, 283–296 (2012).

20. Jones, S. *et al.* Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 4283–4288 (2008).

21. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science (80-.).* **318**, 1108–1113 (2007).

22. Sjöblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science (80-.).* **314**, 268–274 (2006).

23. Olivier, M., Asmis, R., Hawkins, G. A., Howard, T. D. & Cox, L. A. The need for multi-omics biomarker signatures in precision medicine. *International Journal of Molecular Sciences* **20**, 4781 (2019).

24. Schwaederle, M. *et al.* Association of biomarker-based treatment strategies with response rates and progression-free survival in refractory malignant neoplasms a meta-Analysis. *JAMA Oncol.* **2**, 1452–1459 (2016).

25. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. (2014). doi:10.1038/nature12912

26. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* (2011). doi:10.1016/j.cell.2011.02.013

27. Mukherjee, S. Genomics-Guided Immunotherapy for Precision Medicine in Cancer. *Cancer Biotherapy and Radiopharmaceuticals* **34**, 487–497 (2019).

28. Lohmueller, J. & Finn, O. J. Current modalities in cancer immunotherapy: Immunomodulatory antibodies, CARs and vaccines. *Pharmacology and Therapeutics* **178**, 31–47 (2017).

29. Sadreddini, S. *et al.* Immune checkpoint blockade opens a new way to cancer immunotherapy. *Journal of Cellular Physiology* **234**, 8541–8549 (2019).

30. Gomes-Silva, D. & Ramos, C. A. Cancer Immunotherapy Using CAR-T Cells: From the Research Bench to the Assembly Line. *Biotechnology Journal* **13**, (2018).

31. Hegde, P. S. & Chen, D. S. Top 10 Challenges in Cancer Immunotherapy. *Immunity* **52**,

17–35 (2020).

32. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nature Genetics* **45**, 1113–1120 (2013).

33. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).

34. Creixell, P. *et al.* Pathway and network analysis of cancer genomes. *Nature Methods* **12**, 615–621 (2015).

35. Ding, L., Wendl, M. C., McMichael, J. F. & Raphael, B. J. Expanding the computational toolbox for mining cancer genomes. *Nature Reviews Genetics* **15**, 556–570 (2014).

36. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* **6**, 271-281.e7 (2018).

37. Gonzalez-Perez, A., Deu-Pons, J. & Lopez-Bigas, N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med.* **4**, 1–13 (2012).

38. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).

39. Prahallad, A. *et al.* Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature* **483**, 100–104 (2012).

40. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).

41. Heng, H. H. The Genomic Landscape of Cancers. in *Ecology and Evolution of Cancer* 69–86 (Elsevier Inc., 2017). doi:10.1016/B978-0-12-804310-3.00005-3

42. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science (80-.).* **348**, 880–886 (2015).

43. Califano, A. & Alvarez, M. J. The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nat. Rev. Cancer* **17**, 116–130 (2017).

44. Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).

45. Waddington, C. H. Canalization of development and genetic assimilation of acquired characters. *Nature* **183**, 1654–1655 (1959).

46. Young, S. R. *et al.* Establishment and serial passage of cell cultures derived from LuCaP xenografts. **73**, 1251–1262 (2013).

47. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

48. Califano, A., Butte, A. J., Friend, S., Ideker, T. & Schadt, E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nature Genetics* **44**, 841–847 (2012).

49. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).

50. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–50 (2005).

51. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30 (2000).

52. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).

53. Pe'er, D. & Hacohen, N. Principles and strategies for developing network models in cancer. *Cell* **144**, 864–873 (2011).

54. Segal, E. *et al.* Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).

55. Litvin, O., Causton, H. C., Chen, B. J. & Pe'er, D. Modularity and interactions in the genetics of gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 6441–6446 (2009).

56. Mitra, K., Carvunis, A. R., Ramesh, S. K. & Ideker, T. Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* **14**, 719–732 (2013).

57. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. *Discovering regulatory and signalling circuits in molecular interaction networks*. *BIOINFORMATICS* **18**, (2002).

58. Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for detecting significantly mutated pathways in cancer. in *Journal of Computational Biology* **18**, 507–522 (Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA , 2011).

59. Paull, E. O. *et al.* Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). **29**, 2757–2764 (2013).

60. Shi, Y., Inoue, H., Wu, J. C. & Yamanaka, S. Induced pluripotent stem cell technology: A decade of progress. *Nature Reviews Drug Discovery* **16**, 115–130 (2017).

61. Malik, N. & Rao, M. S. A review of the methods for human iPSC derivation. in *Methods in*

Molecular Biology **997**, 23–33 (Humana Press, Totowa, NJ, 2013).

62. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).

63. Basso, K. *et al.* Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* **37**, 382–390 (2005).

64. Carro, M. S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318–325 (2010).

65. Lefebvre, C. *et al.* A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. 377 (2010). doi:10.1038/msb.2010.31

66. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).

67. Anh Huynh-Thu, V. & Sanguinetti, G. *Gene regulatory network inference: an introductory survey*.

68. Wouters, J., Kalender Atak, Z. & Aerts, S. Decoding transcriptional states in cancer. *Current Opinion in Genetics and Development* **43**, 82–92 (2017).

69. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **5**, 12776 (2010).

70. Aibar, S. *et al.* SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

71. Janky, R. *et al.* iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. *PLoS Comput. Biol.* **10**, e1003731 (2014).

72. Yang, J. *et al.* Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell* **117**, 927–939 (2004).

73. Nambu, J. R., Lewis, J. O., Wharton, K. A. & Crews, S. T. The Drosophila single-minded gene encodes a helix-loop-helix protein that acts as a master regulator of CNS midline development. *Cell* **67**, 1157–1167 (1991).

74. Akbani, R. *et al.* Realizing the promise of reverse phase protein arrays for clinical, translational, and basic research: A workshop report the RPPA (Reverse Phase Protein Array) Society. *Mol. Cell. Proteomics* **13**, 1625–1643 (2014).

75. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–47 (2016).

76. Castro, M. A. A. *et al.* Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat. Genet.* **48**, 12–21 (2016).

77. Zhang, S. *et al.* Stroma-associated master regulators of molecular subtypes predict patient prognosis in ovarian cancer. *Sci. Rep.* **5**, 16066 (2015).

78. Yepes, S., Torres, M. M. & López-Kleine, L. Regulatory network reconstruction reveals genes with prognostic value for chronic lymphocytic leukemia. *BMC Genomics* **16**, 1002 (2015).

79. Giorgi, F. M. *et al.* Inferring Protein Modulation from Gene Expression Data Using Conditional Mutual Information. *PLoS One* **9**, e109569 (2014).

80. Quelle, F. W. *et al.* Phosphorylation and activation of the DNA binding activity of purified stat1 by the Janus protein-tyrosine kinases and the epidermal growth factor receptor. *J. Biol. Chem.* **270**, 20775–20780 (1995).

81. Zhao, X. *et al.* The N-Myc-DLL3 Cascade Is Suppressed by the Ubiquitin Ligase Huwe1 to Inhibit Proliferation and Promote Neurogenesis in the Developing Brain. *Dev. Cell* **17**, 210–221 (2009).

82. Wang, K. *et al.* Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nat. Biotechnol.* **27**, 829 (2009).

83. Chen, J. C. *et al.* Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. *Cell* **159**, 402–14 (2014).

84. Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**, 556–560 (2012).

85. Paull, E. O. *et al.* A modular master regulator landscape controls cancer transcriptional identity. *Cell* **184**, 334-351.e20 (2021).

86. Rappoport, N. & Shamir, R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.* **46**, 10546–10562 (2018).

87. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912 (2009).

88. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304. e6 (2018).

89. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, 237–245 (2010).

90. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet Classification with Deep Convolutional Neural Networks*.

91. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, (2018).

92. Chaudhary, K., Poirion, O. B., Lu, L. & Garmire, L. X. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* **24**, 1248–1259 (2018).

93. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

94. Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**, 835-849.e21 (2019).

95. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).

96. Paull, E. O., Jones, S. J., Alvarez, M. & Califano, A. MOMA: Multi Omic Master Regulator Analysis. R package version 1.2.0. *Bioconductor* (2020).

97. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, 1–16 (2004).

98. Paull, E. O. *et al.* MOMA Web App V1.0. (2020). Available at: mr-graph.org.

99. Torres-García, W. *et al.* PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics* **30**, 2224–2226 (2014).

100. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2018).

101. Park, H. S. & Jun, C. H. A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.* **36**, 3336–3341 (2009).

102. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

103. Berglund, E. *et al.* Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.* **9**, 2419 (2018).

104. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462 (2013).

105. Aytes, A. *et al.* Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy. *Cancer*

Cell **25**, 638–651 (2014).

106. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).

107. Hwang, S. *et al.* HumanNet v2: human gene networks for disease research. *Nucleic Acids Res.* **47**, D573–D580 (2018).

108. Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput. Biol.* **9**, e1002886 (2013).

109. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

110. Miyamoto, S., Ichihashi, H. H., Honda, K. & Ichihashi, H. H. *Algorithms for fuzzy clustering*. (Springer, 2008).

111. Drake, J. M. *et al.* Phosphoproteome Integration Reveals Patient-Specific Networks in Prostate Cancer. *Cell* **166**, 1041–1054 (2016).

112. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).

113. Cowley, G. S. *et al.* Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci Data* **1**, 140035 (2014).

114. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812-830.e14 (2018).

115. Tibshirani, R. The lasso method for variable selection in the Cox model. *Stat. Med.* **16**, 385–395 (1997).

116. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

117. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res.* **39**, e118–e118 (2011).

118. Dai, X. *et al.* The ovo gene required for cuticle formation and oogenesis in flies is involved in hair formation and spermatogenesis in mice. *Genes Dev.* **12**, 3452–3463 (1998).

119. Gao, X., Bali, A. S., Randell, S. H. & Hogan, B. L. M. GRHL2 coordinates regeneration of a polarized mucociliary epithelium from basal stem cells. *J. Cell Biol.* **211**, 669–682 (2015).

120. Kappes, D. J. Expanding roles for ThPOK in thymic development. *Immunol. Rev.* **238**, 182–194 (2010).

121. Frisch, S. M., Schaller, M. & Cieply, B. Mechanisms that link the oncogenic epithelial-mesenchymal transition to suppression of anoikis. *Journal of Cell Science* **126**, 21–29 (2013).

122. Jolly, M. K. *et al.* Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget* **7**, 27067–27084 (2016).

123. Handle, F. *et al.* Drivers of AR indifferent anti-androgen resistance in prostate cancer cells. *Sci. Rep.* **9**, (2019).

124. Zhang, D. *et al.* Stem cell and neurogenic gene-expression profiles link prostate basal cells to aggressive prostate cancer. *Nat. Commun.* **7**, (2016).

125. Rajan, P. *et al.* Next-generation sequencing of advanced prostate cancer treated with androgen-deprivation therapy. *Eur. Urol.* **66**, 32–39 (2014).

126. Sun, Y. *et al.* Androgen deprivation causes epithelial-mesenchymal transition in the prostate: Implications for androgen- deprivation therapy. *Cancer Res.* **72**, 527–536 (2012).

127. Tsai, Y. C. *et al.* Androgen deprivation therapy-induced epithelial-mesenchymal transition of prostate cancer through downregulating SPDEF and activating CCL2. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1864**, 1717–1727 (2018).

128. Chuu, C. P. *et al.* Androgens as therapy for androgen receptor-positive castration-resistant prostate cancer. *Journal of Biomedical Science* **18**, 63 (2011).

129. Loeb, S. *et al.* Testosterone replacement therapy and risk of favorable and aggressive prostate cancer. *J. Clin. Oncol.* **35**, 1430–1436 (2017).

130. Alvarez, M. J. *et al.* A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors. *Nat. Genet.* **50**, 979–989 (2018).

131. Baxter, R. J. *Exactly Solved Models in Statistical Mechanics*. (Harcourt Brace Jovanovich Publishers, 1982).

132. Rydenfelt, M., Wongchenko, M., Klinger, B., Yan, Y. & Blüthgen, N. The cancer cell proteome and transcriptome predicts sensitivity to targeted and cytotoxic drugs. *Life Sci. Alliance* **2**, (2019).

133. Kim, J. W. *et al.* Decomposing Oncogenic Transcriptional Signatures to Generate Maps of Divergent Cellular States. *Cell Syst.* **5**, 105-118.e9 (2017).

134. Sankaranarayanan, P., Schomay, T. E., Aiello, K. A. & Alter, O. Tensor GSVD of patient- and platform-matched tumor and normal DNA copy-number profiles uncovers chromosome arm-wide patterns of tumor-exclusive platform-consistent alterations encoding for cell transformation and predicting ovarian cancer survival. *PLoS One* **10**, e0121396 (2015).

135. Malta, T. M. *et al.* Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* **173**, 338-354.e15 (2018).

136. Roig, I. *et al.* Mouse TRIP13/PCH2 is required for recombination and normal higher-order chromosome structure during meiosis. *PLoS Genet.* **6**, e1001062 (2010).

137. Jain, M. *et al.* TOP2A is overexpressed and is a therapeutic target for adrenocortical carcinoma. *Endocr Relat Cancer* **20**, 361–370 (2013).

138. Brosh, R. & Rotter, V. Transcriptional control of the proliferation cluster by the tumor suppressor p53. *Mol Biosyst* **6**, 17–29 (2010).

139. Unoki, M., Brunet, J. & Mousli, M. Drug discovery targeting epigenetic codes: the great potential of UHRF1, which links DNA methylation and histone modifications, as a drug target in cancers and toxoplasmosis. *Biochem Pharmacol* **78**, 1279–1288 (2009).

140. Corpet, A. *et al.* Asf1b, the necessary Asf1 isoform for proliferation, is predictive of outcome in breast cancer. *EMBO J* **30**, 480–493 (2011).

141. Yazawa, T. *et al.* Lack of class II transactivator causes severe deficiency of HLA-DR expression in small cell lung cancer. *J Pathol* **187**, 191–199 (1999).

142. Broyde, J. *et al.* Oncoprotein-specific molecular interaction maps (SigMaps) for cancer network analyses. *Nat. Biotechnol.* **39**, 215–224 (2021).

143. Rajbhandari, P. *et al.* Cross-cohort analysis identifies a TEAD4–MYCN positive feedback loop as the core regulatory element of high-risk neuroblastoma. *Cancer Discov.* **8**, 582–599 (2018).

144. Hu, X. *et al.* TumorFusions: An integrative resource for cancer-associated transcript fusions. *Nucleic Acids Res.* **46**, D1144–D1149 (2018).

145. Lachmann, A., Giorgi, F. M., Lopez, G. & Califano, A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32**, 2233–5 (2016).

146. Giorgi, F. M., Alvarez, M. J. & Califano, A. aracne.networks, a data package containing gene regulatory networks assembled from TCGA data by the ARACNe algorithm. (2016). doi:10.1093/bioinformatics/btw216

147. Jerby-Arnon, L. *et al.* Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell* **158**, 1199–1209 (2014).

148. Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A. & Williams Jr, R. M. *The American Solider : Adjustment During Army Life* . **1**, (Princeton University Press, 1949).

149. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).

150. Repana, D. *et al.* The Network of Cancer Genes (NCG): A comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens 06 Biological Sciences 0604 Genetics 11 Medical and Health Sciences 1112 Oncology and Carcinogenesis 06 Biological Sciences 0601 Biochemistry and Cell Biology. *Genome Biol.* **20**, (2019).

151. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* **69**, 6660–6667 (2009).

152. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

153. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. S. mixtools: An R Package for Analyzing Mixture Models. *2009* **32**, 29 (2009).

154. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003).

155. Kucukural, A., Yukselen, O., Ozata, D. M., Moore, M. J. & Garber, M. DEBrowser: Interactive differential expression analysis and visualization tool for count data 06 Biological Sciences 0604 Genetics 08 Information and Computing Sciences 0806 Information Systems. *BMC Genomics* **20**, (2019).

156. Ding, H. *et al.* Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nat. Commun.* **9**, 1471 (2018).

157. Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).

158. Ciriello, G. *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163**, 506–519 (2015).

159. Robertson, A. G. *et al.* Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell* **171**, 540-556.e25 (2017).

160. Arnold, M., Ferlay, J., Van Berge Henegouwen, M. I. & Soerjomataram, I. Global burden of oesophageal and gastric cancer by histology and subsite in 2018. *Gut* **69**, 1564–1571 (2020).

161. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).

162. Luo, G. *et al.* Global patterns and trends in stomach cancer incidence: Age, period and birth cohort analysis. *Int. J. Cancer* **141**, 1333–1344 (2017).

163. Ku, G. & Ilson, D. Cancer of the Stomach. in *Abeloff's Clinical Oncology* (Elsevier, 2019).

164. Hu, B. *et al.* Gastric cancer: Classification, histology and application of molecular pathology. *Journal of Gastrointestinal Oncology* **3**, 251–261 (2012).

165. Kelley, J. R. & Duggan, J. M. Gastric cancer epidemiology and risk factors. *J. Clin. Epidemiol.* **56**, 1–9 (2003).

166. Miao, R.-L. & Wu, A.-W. Towards personalized perioperative treatment for advanced gastric cancer. *World J. Gastroenterol.* **20**, 11586–94 (2014).

167. Watanabe, M. *et al.* Recent progress in multidisciplinary treatment for patients with esophageal cancer. *Surgery Today* **50**, 12–20 (2020).

168. Biagioni, A. *et al.* Update on gastric cancer treatments and gene therapies. *Cancer and Metastasis Reviews* **38**, 537–548 (2019).

169. GASTRIC (Global Advanced/Adjuvant Stomach Tumor Research International Collaboration) Group *et al.* Benefit of adjuvant chemotherapy for resectable gastric cancer: a meta-analysis. *JAMA* **303**, 1729–1737 (2010).

170. Hermans, J. *et al.* Adjuvant therapy after curative resection for gastric cancer: meta-analysis of randomized trials. *J. Clin. Oncol.* **11**, 1441–7 (1993).

171. Janunger, K.-G., Hafström, L. & Glimelius, B. Chemotherapy in gastric cancer: a review and updated meta-analysis. *Eur. J. Surg.* **168**, 597–608 (2002).

172. Bang, Y.-J. *et al.* Adjuvant capecitabine and oxaliplatin for gastric cancer after D2 gastrectomy (CLASSIC): a phase 3 open-label, randomised controlled trial. *Lancet (London, England)* **379**, 315–21 (2012).

173. Ajani, J. A. *et al.* Multicenter phase III comparison of cisplatin/S-1 with cisplatin/infusional fluorouracil in advanced gastric or gastroesophageal adenocarcinoma study: the FLAGS trial. *J. Clin. Oncol.* **28**, 1547–53 (2010).

174. Sakuramoto, S. *et al.* Adjuvant chemotherapy for gastric cancer with S-1, an oral fluoropyrimidine. *N. Engl. J. Med.* **357**, 1810–20 (2007).

175. Salati, M. *et al.* Gastric cancer: Translating novels concepts into clinical practice. *Cancer Treatment Reviews* **79**, 101889 (2019).

176. Kurokawa, Y. *et al.* Multicenter large-scale study of prognostic impact of HER2 expression in patients with resectable gastric cancer. *Gastric Cancer* **18**, 691–697 (2015).

177. Sukawa, Y. *et al.* HER2 Expression and PI3K-Akt Pathway Alterations in Gastric Cancer Significance of HER2 Expression in Gastric Cancer. *Digestion* **89**, 12–17 (2014).

178. Kelly, R. J. *et al.* Adjuvant Nivolumab in Resected Esophageal or Gastroesophageal Junction Cancer. *N. Engl. J. Med.* **384**, 1191–1203 (2021).

179. Hecht, J. R. *et al.* Lapatinib in combination with capecitabine plus oxaliplatin in human epidermal growth factor receptor 2-positive advanced or metastatic gastric, esophageal, or gastroesophageal adenocarcinoma: TRIO-013/LOGiC - A randomized phase III trial. *J. Clin. Oncol.* **34**, 443–451 (2016).

180. Lordick, F. *et al.* Capecitabine and cisplatin with or without cetuximab for patients with previously untreated advanced gastric cancer (EXPAND): A randomised, open-label phase 3 trial. *Lancet Oncol.* **14**, 490–499 (2013).

181. Waddell, T. *et al.* Epirubicin, oxaliplatin, and capecitabine with or without panitumumab for patients with previously untreated advanced oesophagogastric cancer (REAL3): A randomised, open-label phase 3 trial. *Lancet Oncol.* **14**, 481–489 (2013).

182. Catenacci, D. V. T. *et al.* Rilotumumab plus epirubicin, cisplatin, and capecitabine as first-line therapy in advanced MET-positive gastric or gastro-oesophageal junction cancer (RILOMET-1): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol.* **18**, 1467–1482 (2017).

183. Shah, M. A. *et al.* Effect of fluorouracil, leucovorin, and oxaliplatin with or without onartuzumab in HER2-negative, MET-positive gastroesophageal adenocarcinoma: The METGastric randomized clinical trial. *JAMA Oncol.* **3**, 620–627 (2017).

184. Ohtsu, A. *et al.* Everolimus for previously treated advanced gastric cancer: Results of the randomized, double-blind, phase III GRANITE-1 study. *J. Clin. Oncol.* **31**, 3935–3943 (2013).

185. Bang, Y. J. *et al.* Olaparib in combination with paclitaxel in patients with advanced gastric cancer who have progressed following first-line therapy (GOLD): a double-blind, randomised, placebo-controlled, phase 3 trial. *Lancet Oncol.* **18**, 1637–1651 (2017).

186. Lei, Z. *et al.* Identification of Molecular Subtypes of Gastric Cancer With Different Responses to PI3-Kinase Inhibitors and 5-Fluorouracil. *Gastroenterology* **145**, 554–565 (2013).

187. Yong, W. P. *et al.* Real-time tumor gene expression profiling to direct gastric cancer chemotherapy: Proof-of-concept '3G' trial. *Clin. Cancer Res.* **24**, 5272–5281 (2018).

188. Bass, A. J. *et al.* Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).

189. Liu, Y. *et al.* Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. *Cancer Cell* **33**, 721-735.e8 (2018).

190. Sohn, B. H. *et al.* Clinical Significance of Four Molecular Subtypes of Gastric Cancer Identified by The Cancer Genome Atlas Project. *Clin. Cancer Res.* **23**, 4441–4449 (2017).

191. Pietrantonio, F. *et al.* MSI-GC-01: Individual patient data (IPD) meta-analysis of microsatellite instability (MSI) and gastric cancer (GC) from four randomized clinical trials (RCTs). *J. Clin. Oncol.* **37**, 66–66 (2019).

192. Wong, S. S. *et al.* Genomic landscape and genetic heterogeneity in gastric adenocarcinoma revealed by whole-genome sequencing. *Nat. Commun.* **5**, 5477 (2014).

193. Lee, J. *et al.* Nanostring-Based Multigene Assay to Predict Recurrence for Gastric Cancer Patients after Surgery. *PLoS One* **9**, e90133 (2014).

194. Cristescu, R. *et al.* Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nat. Med.* **21**, 449–456 (2015).

195. Ding, H., Wang, W. & Califano, A. iterClust: a statistical framework for iterative clustering analysis. *Bioinformatics* **34**, 2865–2866 (2018).

196. Miranda, A. *et al.* Cancer stemness, intratumoral heterogeneity, and immune response across cancers. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 9020–9029 (2019).

197. Liu, X., Zhang, Z. & Zhao, G. Recent advances in the study of regulatory T cells in gastric cancer. *International Immunopharmacology* **73**, 560–567 (2019).

198. Seeler, J.-S. & Dejean, A. SUMO and the robustness of cancer. *Nat. Rev. Cancer* **17**, 184–197 (2017).

199. Nakata, S., Fujita, M. & Nakanishi, H. Efficacy of Afatinib and Lapatinib against HER2 Gene-amplified Trastuzumab-sensitive and -resistant Human Gastric Cancer Cells. *Anticancer Res.* **39**, 5927–5932 (2019).

200. Blackwell, K. L. *et al.* Randomized study of lapatinib alone or in combination with trastuzumab in women with ErbB2-positive, trastuzumab-refractory metastatic breast cancer. *J. Clin. Oncol.* **28**, 1124–1130 (2010).

201. Satoh, T. *et al.* Lapatinib plus paclitaxel versus paclitaxel alone in the second-line treatment of HER2-amplified advanced gastric cancer in Asian populations: TyTAN - A randomized, phase III study. *J. Clin. Oncol.* **32**, 2039–2049 (2014).

202. Chen, Z. *et al.* Characterization and validation of potential therapeutic targets based on the molecular signature of patient-derived xenografts in gastric cancer. *J. Hematol. Oncol.* **11**, 20 (2018).

203. Yoshioka, T. *et al.* Antitumor activity of pan-HER inhibitors in HER2-positive gastric cancer. (2018). doi:10.1111/cas.13546

204. Kursa, M. B. & Rudnicki, W. R. Feature selection with the boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).

Appendix A: Supplement for "A modular master regulator landscape controls cancer transcriptional identity"

(Note: Only supplementary figures from the paper are provided. Due to their size and complexity supplementary table files were not included but they can be downloaded from the online version of the publication or at mr-graph.org).

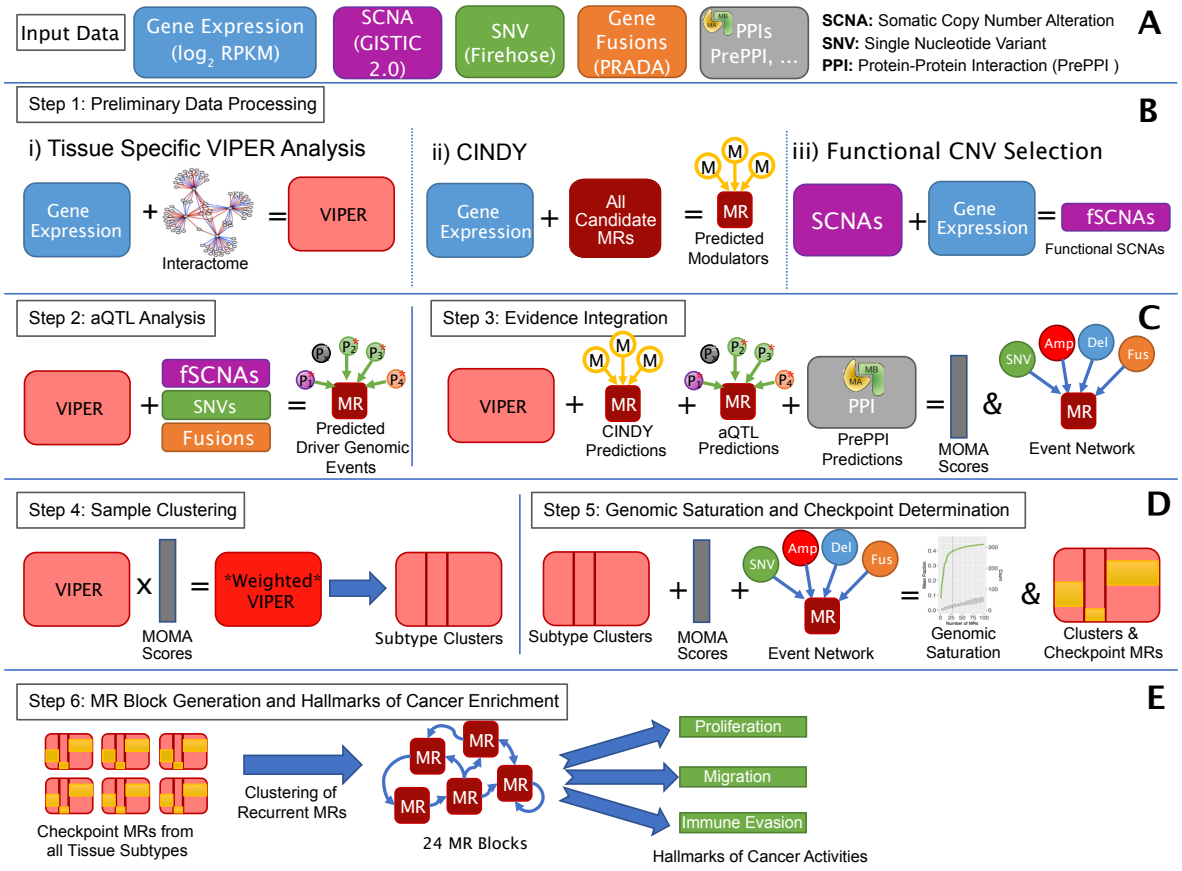


Figure S2.1 Detailed Conceptual Flowchart of MOMA.
(A) Input data for the MOMA pipeline. Data types are coded by color consistently throughout the manuscript. **(B)** (i) VIPER inference of protein activity from Gene Expression Profile data for each sample. Tumor-specific ARACNe networks are used for the analysis. (ii) CINDy modulator predictions based on conditional mutual information analysis from gene expression profiles. (iii) Selection of functional SCNA (fSCNA) by measuring the statistical independence of gene copy number and expression, thus removing the vast majority of candidate SCNA genes. **(C)** Statistical significance from CINDy (i.e., prediction of genes upstream of one or more MR proteins), aQTL analysis (i.e., prediction of genes whose genetic alteration is associated with differential MR activity), and PrePPI (i.e., genes encoding for proteins that physically interact with one or more MRs) is integrated using Fisher's method. This produces a tumor cohort-specific *MOMA score* and rank for each candidate MR protein that integrates both gene expression and mutational profile information. The analysis also associates each candidate MR with a specific set of recurrent genomic alterations in its upstream pathways, the "event network." Finally, **(D)** for each cohort, VIPER-inferred protein activity vectors, weighted by the corresponding MOMA score vectors, are clustered to identify molecularly distinct tumor subtypes. This is followed by genomic saturation analysis on a per-subtype/per-sample basis to refine the MR repertoire comprising each subtype's Tumor Checkpoint. **(E)** Finally, cluster analysis of recurrent Tumor Checkpoint MRs is performed to reveal highly recurrent sub-modular MR structures (MR-Blocks/MRBs), each one regulating a unique set of Cancer Hallmarks.

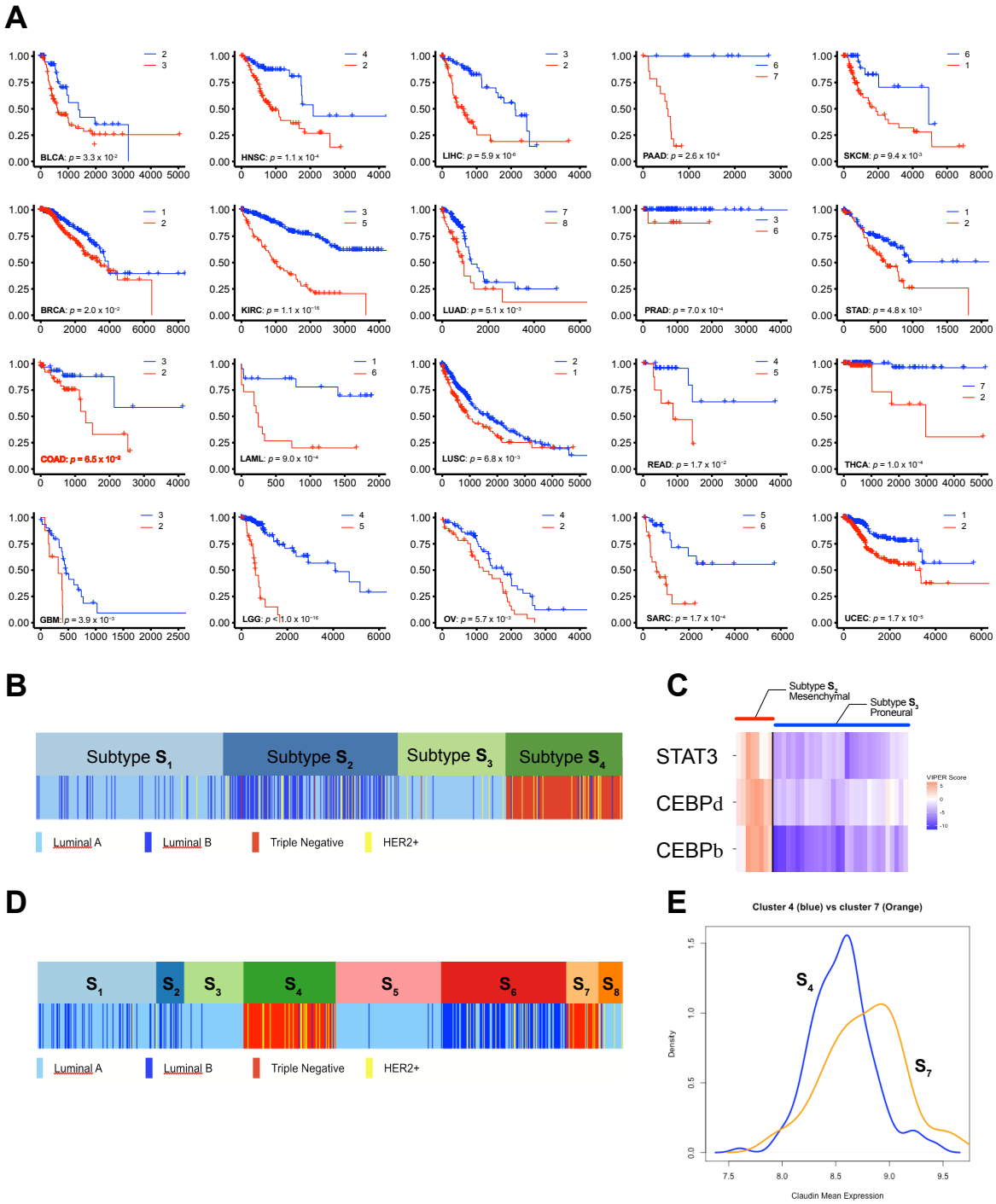


Figure S2.2 Functional validation of MOMA subtypes and survival segregation.
(A) Kaplan-Meyer survival plots for the best and worst-outcome subtype, for each of the 20 TCGA cohorts. Survival time in days and survival probability are shown on the x- and y-axes, respectively. *P*-values for the COX proportional hazard model test between subtypes are reported in each plot, with non-significant ones (i.e., COAD) shown in red. Subtype Ids are shown in the legend of each plot. **(B)** Overlap of MOMA-inferred BRCA subtypes (S_1 – S_4) with classical

subtypes—Luminal A, Luminal B, Basal, and HER2+. As shown, S_1 and S_3 are highly enriched in Luminal A samples, S_2 in Luminal B samples, and S_4 in Basal samples. **(C)** VIPER-inferred protein activity heatmap for STAT3, CEBPδ and CEBPβ in subtype S_2 (mesenchymal) vs. S_3 (proneural) of the GBM cohort. Consistent with previous publications, showing their role as synergistic MRs of the mesenchymal subtype of GBM[64], these proteins are aberrantly co-activated in S_2 but not in S_3. **(D – E)** An 8-cluster solution in BRCA splits the triple negative tumors into two smaller subtypes associated with low and high Claudin expression, respectively.

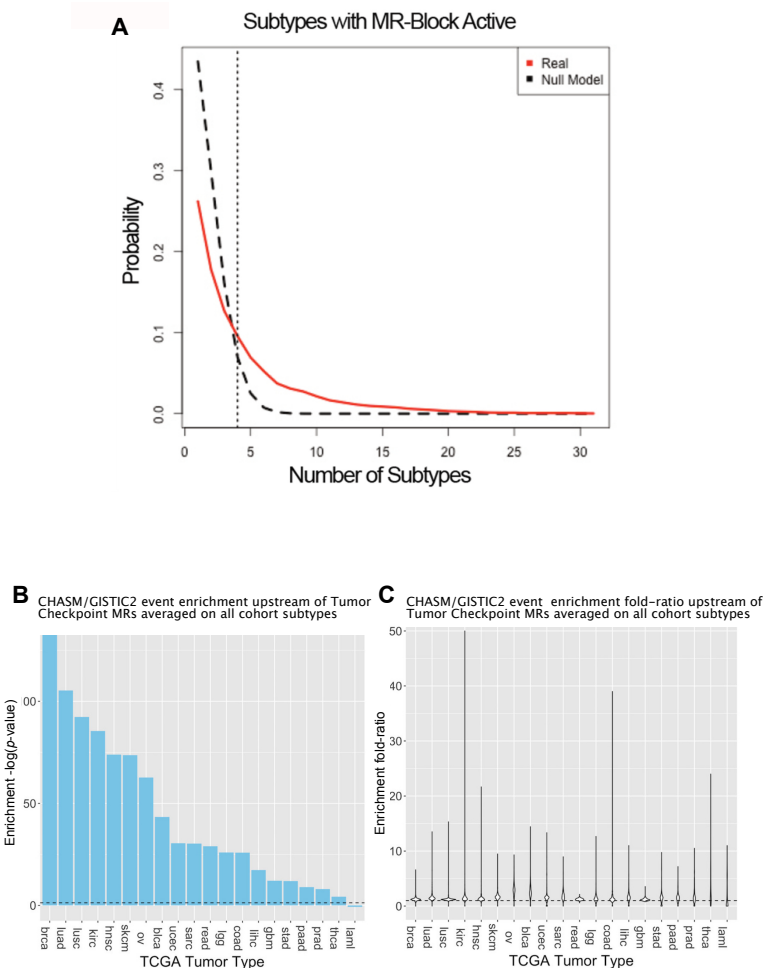


Figure S2.3 Checkpoint proteins are highly recurrent and downstream of driver genomic events

(A) The statistical significance of MR recurrence in $\geq k$ subtypes was assessed using a null hypothesis based on 100 random selections of an identical number of proteins, for each Tumor Checkpoint, from all possible regulatory proteins ($N = 2{,}506$) (black curve). From this analysis $k \geq 4$ (vertical dotted line) emerged as an appropriate threshold for statistical significance ($p < 0.05$). The actual distribution of MOMA-inferred MR recurrence in k Tumor Checkpoints is shown as a red curve. The difference between the real and null-hypothesis distributions is highly statistically significantly ($p < 2.2\times10^{-16}$, by non-parametric Kolmogorov–Smirnov test), indicating that MOMA-inferred MRs are highly recurrent across multiple tumor subtypes. **(B)** Statistical significance of genomic driver gene enrichment upstream of predicted Tumor Checkpoint MRs, in each tumor cohort—including single nucleotide somatic variants detected by CHASM and focal copy number variants detected by GISTIC 2.0. -Log_{10} p-values are shown as bar plots, with a horizontal dashed line representing the statistical significance threshold ($p = 0.05$). **(C)** Violin plots represent the enrichment ratio probability densities for CHASM/GISTIC2.0 events vs. all mutational events, upstream of predicted Tumor Checkpoint MRs, for each tissue type, on a sample by sample basis. This suggests mutations identified upstream of Tumor Checkpoint MRs are much more likely to be drivers than passengers.

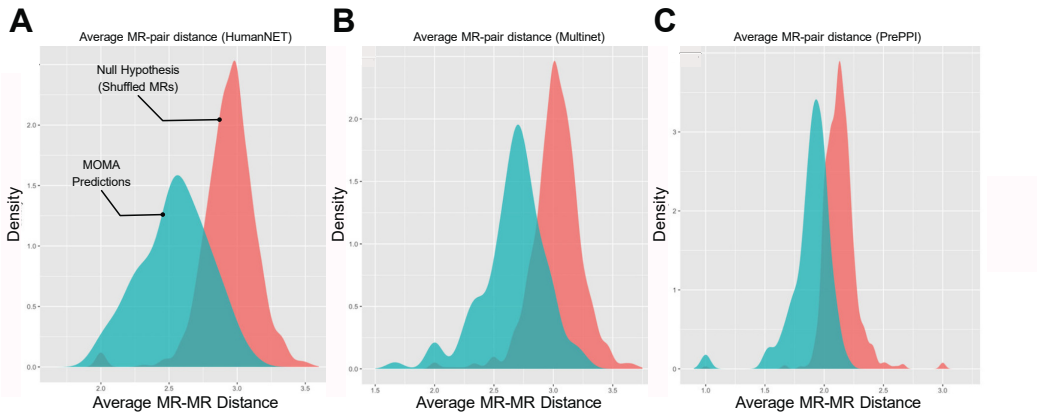


Figure S2.4 Recurrent MRs are predicted to be hyperconnected and modular.
Probability density plots of the mean shortest path distance between all predicted MR pairs, in each Tumor Checkpoint (blue), compared with pairwise distances between random regulatory protein pairs, in **(A)** the HumanNet network **(B)** the Multinet network, and **(C)** the PrePPI protein-protein interaction network.

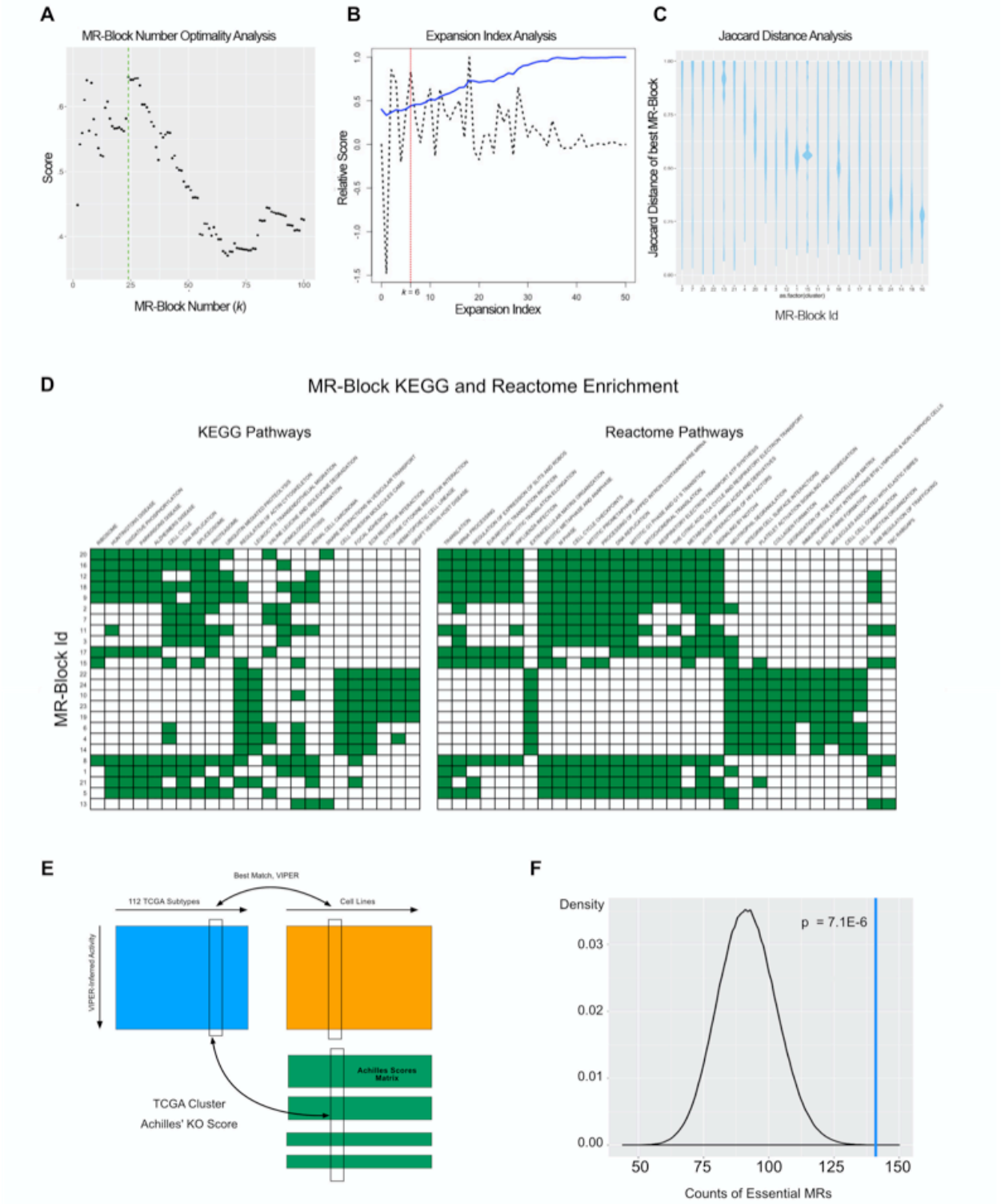


Figure S2.5 MR-Block (MRB) cluster analysis, Cancer Hallmark enrichment analysis, and Achilles' essentiality analysis

(A) The analytical clustering score for the 407 recurrent MR proteins, across all tissue types, for $k = 2$ to 100 clusters, was used to identify $k = 24$ as the optimal MRB number (green line). **(B)** Relative score representing the enrichment specificity of the 24 MRBs in tumor hallmarks proteins

(y-axis) as an additional *k* MRs are added to each MRB by the fuzzy-clustering analysis (x-axis). The blue curve represents the Eigen-trace of the covariance matrix of all hallmark enrichments for all MRBs (MRB specificity), while the dashed black line represents the delta enrichment (i.e., increase or decrease) with respect to the *k* - 1 solution. We selected *k* = 6 (dotted red line) as the optimal fuzzy-clustering expansion index as it represents the smallest absolute maximum with one of the highest specificity increase. **(C)** Violin plots of the Jaccard concordance index (MR overlap) for each of the 24 MRBs with the most similar MRB identified by every other clustering solution (*k* = 2 to 100, excluding *k* = 24). MRBs are sorted left to right, from the most conserved across all clustering solutions (MRB:2) to the one most unique to the k = 24 optimal solution (MRB:16). The latter is still consistent with MRBs in >25% of the clustering solutions, confirming the analysis robustness. **(D)** Enrichment of MRB MR target genes in KEGG and Reactome gene sets (FDR < 0.05). Given the much larger set of gene sets in these databases, only the top 3 most significant for each MRB are shown. More than 3 may be shown when statistical ties are determined for 1st, 2nd or 3rd place. **(E)** Conceptual representation of the cross-comparison between MOMA-inferred MRs and Achilles-based essential genes. For each MOMA subtype the cell lines that best recapitulate its MR activity signature are identified comparison of VIPER inferred protein activity profiles ($p < 0.01$, Bonferroni corrected, see STAR Methods). This is based on the enrichment of subtype checkpoint MRs in proteins that are differentially active in each cell line. The Achilles's K.O. score of matching cell lines is then averaged to assess overall MR essentiality. Finally, MRB Enrichment in essential MRs is computed by Fisher's Exact Test. **(F)** MR protein overlap with Achilles's significantly essential genes (Bonferroni corrected $p \leq 10^{-5}$) across all MOMA-inferred checkpoint MRs (n = 141, blue vertical line), compared with the probability density generated by 10^6 random selections of the same number regulatory proteins for each subtype and fitted to a normal distribution to asses statistical significance (black curve) ($p < 7.1 \times 10^{-6}$). This is highly conservative as many checkpoint MRs control phenotypes such as EMT, inflammation, apical junction, etc., that cannot be assessed by viability assays *in vitro*.

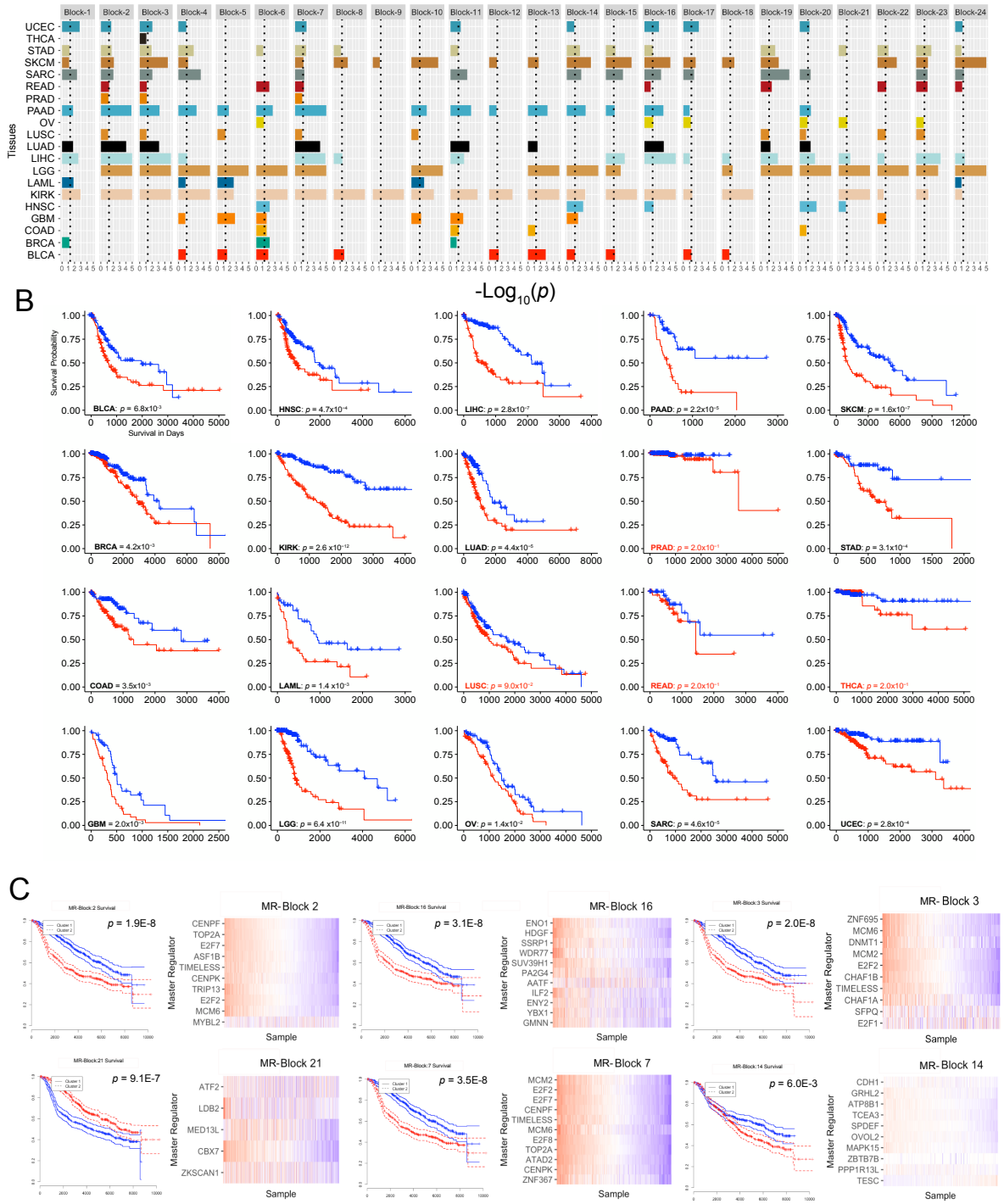


Figure S2.6 Survival stratification by MRB activity

(A) The statistical significance of single-variable Cox regression models for patient survival in each TCGA cohort (*rows*) is shown for each MRB (*columns*). Each colored bar represents the -Log10(*p*) significance of the MRB-based predictor. Bars are truncated at -Log10(*p*) = 5 to improve visualization; non-statistically significant values are omitted. Dashed lines represent the statistical significance threshold ($p = 0.05$). **(B)** Survival plots of all 20 TCGA cohorts using a regularized Cox proportional hazards model trained on the activity of the 24 MRBs. The activity

of an MRB is computed as the average activity of its MRs. The statistical significance of each cox regression model (coefficient), which can include multiple MRBs, is below each plot, with the non-significant cohorts colored in red. Censored marks are shown along each curve. **(C)** Analysis of the 7 MRBs contributing to survival stratification of TCGA BRCA samples in the METABRIC breast cancer dataset (Figure 2.5C). Differential survival is shown for METABRIC samples with positive (red) vs. negative (blue) activity of MRB MRs. MRs lacking an ARACNe regulon in the METABRIC analysis were omitted. Survival stratification was most significant for MRB:2, 3, 7, 16, and 21 ($p_2 = 1.9\times10^{-8}$, $p_3 = 2.0\times10^{-8}$, $p_7 = 3.5\times10^{-8}$, $p_{16} = 3.1\times10^{-8}$, $p_{21} = 9.1\times10^{-7}$, MRB:11 not shown as not statistically significant).

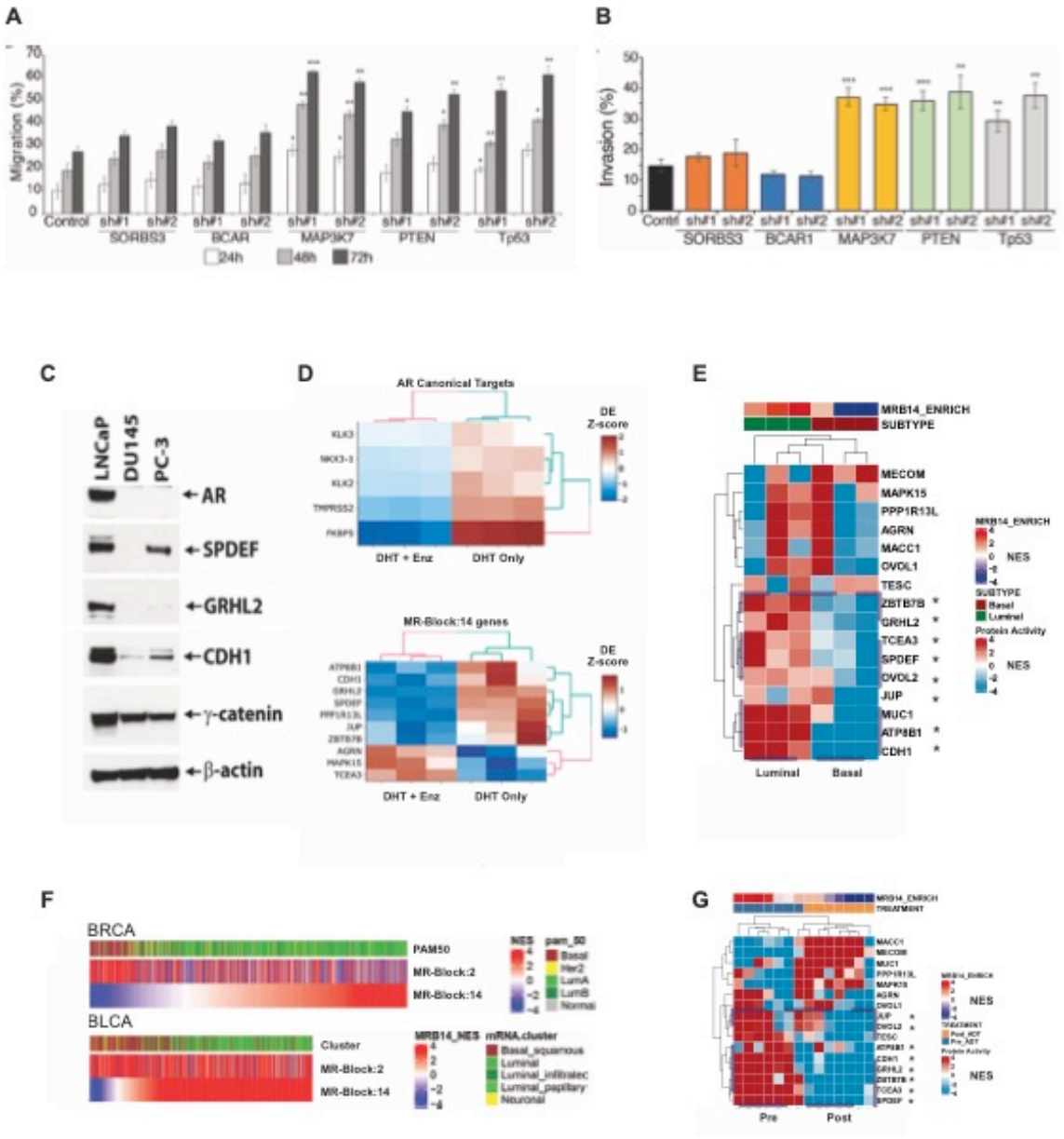


Figure S2.7 MRB:2 and MRB:14 Validation

(A) Results of wound healing assays following shRNA mediated silencing of genes harboring the top 5 mutations upstream of MRB:2, for each of two distinct hairpins per gene, see Figure 2.7C,D for integrated results. **(B)** Results of Boyden Chamber assays following shRNA-mediated silencing of genes harboring the top 5 mutations upstream of MRB:2, for each of two distinct hairpins per gene, see Figure 2.7E for integrated results. **(C)** Western blot assays of LNCaP, DU145 and PC-3 prostate cancer cells confirmed that MRB:14 activity tracks with AR signals. Key MR proteins—SPDEF, GRHL2, CDH1 and γ-catenin—were expressed in LNCaP cells (AR sensitive) and suppressed in AR-independent cells (PC3 and DU145). **(D)** This was further confirmed by Differential Gene Expression (DGE) analysis of MRB:14 MRs in LNCaP cells grown in the presence of androgen (*DHT*) and treated with or without enzalutamide (GEO accession GSE130534). Top panel shows differential expression heatmap of select canonical AR gene targets

(Benjamini-Hochberg FDR ≤ 0.05; fold-change FC ≥ 1.5). Lower panel shows genes encoding for MRB:14 proteins that passed cutoff (FDR ≤ 0.1; FC ≥ 1.2). **(E)** Segregation of normal prostate luminal and basal compartments based on MRB:14 protein activity as inferred by VIPER. The heatmap shows hierarchical clustering of VIPER-inferred activity of MRB:14 MRs, across prostate-derived purified luminal and basal epithelial cells, in triplicate, based on RNA-Seq profile analysis (GEO dataset accession GSE067070). Inconsistent differential expression vs. activity behavior for AGRN, MAPK15, and TCEA3 is consistent with the reported 20% – 30% rate of MR-activity inversion by VIPER where absolute level of activity is correctly computed but the sign (i.e., positive vs. negative MR) is inverted due to autoregulatory loops **(F)** MRB:14 activity tracks with luminal vs. basal classification—based on PAM-50 gene signature analysis—in breast (BRCA) and bladder (BLCA) TCGA cohorts. **(G)** MRB:14 MR activity is inverted in patients undergoing androgen deprivation therapy (ADT), as shown by VIPER-based analysis of RNA-Seq data from 7 TRUSS patient-matched biopsy pairs, pre- and post-ADT (GEO accession GSE48403). Hierarchical clustering using MRB:14 proteins activity provides complete segregation of the two phenotypes.